UNIVERSITY OF
LIVERPOOL

LOW POWER SMALL GEOMETRY BUILDING BLOCKS FOR NEURAL NETWORKS BASED ON CHARGE TRANSFER DEVICES

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy

by

Yajie Chen

November 2008

ABSTRACT

The fast progress in biological study has attracted a growing interest in mimicking the signal processing functions of biological neural systems, which can be implemented in hardware and used to inspire new techniques for real time computations. On the other hand, the ITRS roadmap for Si indicates that alternative paradigms for building computational machines will be required within about 10 years and the massive parallelism of neural systems represents an attractive option. However most implementation approaches are restrictive in physical dimensions and characteristics towards biological networks in hardware. Thus there is a pressing need for compact, low power neural building blocks with operational characteristics that closely mimic realistic neuron cells.

In this thesis, a charge coupled synapse is developed as a core building block for spiking neural networks in hardware. The proposed silicon synapse is based on a two-phase charge transfer device with associated localized memory capability. The correspondence between the fundamental semiconductor processes and the required biological functionality is illustrated by theory. The basic principles of the charge coupled synapse are demonstrated to exhibit a spiking characteristic of biological synapses. Issues related to the analytical modeling of the charge transfer and associated spike generation are investigated by using the Silvaco software package. The device simulations prove that the proposed synaptic device is able to capture the intrinsic dynamics of the biological synapses in spiking neural networks.

This thesis then proposes the device concepts for the programmable dynamic charge coupled synapses where the recovery process of weight charge packet can be greatly accelerated. The first programmable dynamic synapse is based on the charge coupled synapse to which an additional source of minority carriers is attached. The programmable functionality of the synapse is implemented by the weight restoration through charge injection from an N+ implant pulsed by a small negative voltage. An alternative type of programmable dynamic charge coupled synapse consists of an injector MOS transistor in proximity to two MOS capacitors. This approach has the advantage that it can be fabricated through a standard CMOS mixed-signal process.

The function of the MOS transistor is to restore the charge in the well where the duration of this process is dictated by the associated gate voltage. Therefore, the synapse is capable of operating in the facilitating state over a large frequency range. Simulation results are presented which clearly demonstrate its operation.

A low power compact neuron cell in hardware that accommodates charge coupled synapses is designed. The fundamental functionality of the biological neuron cell is implemented by current mirror summing, charge integration onto a thresholding inverter and subsequent slow leakage of charge via a reverse biased diode. The mathematical analysis and Spice simulation are presented to support this work.

By using the innate properties of semiconductors, a charge coupled synapse capable of mimicking spiking dynamics and synaptic plasticity, serves as the compact component for implementing neural networks in silicon, and since it operates in transient mode, its power consumption is negligible. The solid-state neuron presented, performing the integration of synaptic signals, captures the time dependency of the post-synaptic membrane decay, and fires when a certain threshold is reached. This silicon neuron cell together with an array of synapses will provide, for the first time, a core building block that is not only biologically plausible but has the potential to significantly advance the hardware implementation of spiking neural networks towards the biological-scale, using well proven and robust silicon technology.

ACKNOWLEDGEMENTS

First and foremost, I wish to express my sincere gratitude to my supervisor, Prof. Steve Hall, for the things he taught me and the confidence he brought to me over the years. His guidance, encouragement and critical comments have greatly contributed to my research and to the preparation of this thesis. I was inspired by his knowledge, wisdom, and more importantly, by his attitude for research, which will be a treasure of my lifetime.

I would like to thank Dr. O. Buiu and Dr. L. McDaid for their helpful discussions, constructive criticism, and collaboration on this work. They always patiently answer my questions and share their insights on hardware implementation of neural based computational systems.

I would also like to thank all members of the Solid-state Electronics group, especially Dr. Y. Lu, Mr. T. Dowrick, Mr. L. Tan, for helping me along on my way.

Additionally, I wish to thank Prof. Q. H. Wu, Prof. Z. Ji and Mr. L. Jiang for their encouragement and help. Also many thanks to all my friends in Liverpool for their friendship and support.

Last, but by no means least, I am deeply grateful to my family for their encouragement and love, and my girlfriend Dongsong Li for her love, support and belief during my Ph.D study. I will never be able to adequately thank my mother Ms. J. Lai, for the strength, guidance, and love she has given all the years of my life.

CONTENTS

CHAPTER 1 INTRODUCTION 1

SECTION 1.1 TOWARDS INTELLIGENT COMPUTATION SYSTEM 1

SECTION 1.2 NEURAL NETWORKS 4

Section 1.2.1 Biological Neural Networks 4

Section 1.2.2 Artificial Neural Networks 6

Section 1.2.3 Spiking Neural Networks 8

Section 1.2.4 Spiking Neuron Models 9

Section 1.2.5 Synaptic Plasticity 12

SECTION 1.3 NEURAL NETWORKS IN VLSI 13

SECTION 1.4 SILICON NEURON 16

SECTION 1.5 SILICON SYNAPSE 18

SECTION 1.6 OVERVIEW OF THE PROJECT 21

SECTION 1.7 ORGANIZATION OF THE THESIS 24

REFERENCES 25

CHAPTER 2 FUNDAMENTALS OF SEMICONDUCTOR DEVICES 34

SECTION 2.1 INTRODUCTION 34

SECTION 2.2 MOS CAPACITOR 35

Section 2.2.1 Basic Operations 36

Section 2.2.2 Capacitance-Voltage Characteristics 40

Section 2.2.3 Deep Depletion Capacitance 42

Section 2.2.4 C-V Measurement and Parameter Extraction 43

SECTION 2.3 MOS TRANSISTOR 46

Section 2.3.1 I-V Characteristics 47

Section 2.3.2 Linear Operation 48

Section 2.3.3 Saturation Operation 49

Section 2.3.4 Subthreshold Operation 50

SECTION 2.4 CONCLUSIONS 51

REFERENCES 51

CHAPTER 3 CHARGE COUPLED SYNAPSE 53

SECTION 3.1 INTRODUCTION 53
SECTION 3.2 CHARGE COUPLED SYNAPSE MODEL 54
SECTION 3.3 VOLTAGE-DEPENDENT SYNAPTIC WEIGHT 56
SECTION 3.4 SYNAPTIC WEIGHTING PROCESS 58
Section 3.4.1 Self-induced Drift *61*
Section 3.4.2 Thermal Diffusion *63*
Section 3.4.3 Fringing Field Drift *64*
SECTION 3.5 FLOATING DIFFUSION OUTPUT STAGE 65
SECTION 3.6 SIMULATION STUDY AND RESULTS ANALYSIS 70
Section 3.6.1 Synaptic Weight Generation *71*
Section 3.6.2 Synaptic Weighting Process *72*
Section 3.6.3 Spiking Output Current *80*
Section 3.6.4 Post-Synaptic Potential *85*
SECTION 3.7 DISCUSSION AND CONCLUSIONS 88
REFERENCES 90

CHAPTER 4 PROGRAMMABLE CHARGE COUPLED SYNAPSE 91

SECTION 4.1 INTRODUCTION 91
SECTION 4.2 TRANSIENT OPERATION OF WEIGHT MOS CAPACITOR 94
SECTION 4.3 TRANSIENT RESPONSE TO SUCCESSIVE PRE-SYNAPTIC SPIKES 100
SECTION 4.4 PROGRAMMABLE SYNAPSE WITH CHARGE INJECTOR 102
SECTION 4.5 PROGRAMMABLE SYNAPSE WITH INJECTOR TRANSISTOR 106
Section 4.5.1 Device Operation *106*
Section 4.5.2 Analysis and Simulation Results *108*
SECTION 4.6 DISCUSSION AND CONCLUSIONS 113
REFERENCES 114

CHAPTER 5 SILICON NEURON STANDARD CELL 116

SECTION 5.1 INTRODUCTION 116
SECTION 5.2 ANALOG NEURON CELL CIRCUIT 118
Section 5.2.1 Synapses-Neuron Interfacing *119*
Section 5.2.2 Current Mirror Operation *121*
Section 5.2.3 Membrane Potential Generation *122*
Section 5.2.4 Thresholding Operation *125*

SECTION 5.3 SIMULATION STUDY OF NEURON CELL CIRCUIT 125
Section 5.3.1 Synchronous Signal Response *127*
Section 5.3.2 Asynchronous Signal Response *127*
SECTION 5.4 INTEGRATION OF SYNAPSES WITH CHARGE INJECTOR 131
Section 5.4.1 Synchronous Synaptic Signal *133*
Section 5.4.2 Asynchronous Synaptic Signal *135*
SECTION 5.5 INTEGRATION OF SYNAPSES WITH INJECTOR TRANSISTOR 138
SECTION 5.6 NEURON MOS TRANSISTOR 143
Section 5.6.1 neuMOS Principles *143*
Section 5.6.2 Interconnecting Regime *145*
SECTION 5.7 DISCUSSION AND CONCLUSIONS 147
REFERENCES 147
CHAPTER 6 CONCLUSIONS AND FUTURE WORK 149
APPENDIX 1 DESCRIPTION OF SPICE MODEL PARAMETERS 153
APPENDIX 2 ASSOCIATED PUBLICATIONS 154

LIST OF SYMBOLS

Symbol	Description	Unit
A	Area	cm^2
C	Capacitance	F (F/cm^2)
C_d	Depletion layer capacitance per unit area	F/cm^2
C_{dm}	Maximum depletion layer capacitance per unit area	F/cm^2
C_{FN}	Floating diffusion node capacitance	F
C_{gs}	Gate-source capacitance	F
C_{ox}	Oxide capacitance per unit area	F/cm^2
C_{ov}	Overlap capacitance	F
C_{si}	Silicon capacitance per unit area	F/cm^2
C_T	Total capacitance associated with floating gate	F
D	Diffusion coefficient	cm^2/s
D_{eff}	Self-induced drift coefficient	cm^2/s
E	Energy	eV
E_c	Conduction band edge	eV
E_F	Fermi energy level	eV
E_{Fi}	Intrinsic Fermi level	eV
E_g	Energy bandgap	eV
E_v	Valence band edge	eV
E_{vacuum}	Vacuum level	eV
ξ_F	Fringing field	V/cm
ξ_{ox}	Electric field in oxide	V/cm
ξ_S	Self-induced field	V/cm
ξ_{si}	Electric field in silicon	V/cm
ε_0	Vacuum permittivity (=8.85×10^{-14} F/cm)	F/cm
ε_{ox}	Oxide dielectric constant (=3.9)	
ε_{si}	Silicon dielectric constant (=11.9)	
Φ_m	Metal work function	eV
Φ_{si}	Silicon work function	eV
ϕ_B	Bulk potential	V

ϕ_s	Surface potential	V
η	Charge transfer efficiency	
I	Current	A
I_0	Off-current of MOS transistor	A
I_{ds}	Drain-source current	A
I_{leak}	Leaky current	A
I_{sat}	Saturation current	A
I_{sub}	Subthreshold current	A
i_{sum}	Mirrored current	A
J	Current density	A/cm^2
k	Boltzmann's constant (=1.38×10^{-23} J/K)	J/K
μ	Carrier mobility	$cm^2/V\cdot s$
μ_n	Electron mobility	$cm^2/V\cdot s$
μ_p	Hole mobility	$cm^2/V\cdot s$
L	Channel (electrode) length	cm
L_{FN}	Length of floating diffusion node	cm
L_{ov}	Overlap length	cm
λ	Channel length modulation factor	V^{-1}
m	Gate channel coupling	
n	Number of synapses	
n_i	Intrinsic carrier density (=1.45×10^{10} cm^{-3})	cm^{-3}
n_{inv}	Charge density in inversion layer	cm^{-2}
n_{ox}	Charge density in oxide	cm^{-2}
n_p	Injected electron concentration	cm^{-3}
n_W	Synaptic weight charge density per unit area	cm^{-2}
N_a	Acceptor impurity density	cm^{-3}
N_d	Donor impurity density	cm^{-3}
q	Electronic charge (=1.6×10^{-19} C)	C
Q	Charge	C
Q_d	Charge per unit area in the depletion layer	C/cm^2
Q_f	Charge per unit area remaining in storage well	C/cm^2
Q_{inv}	Charge per unit area in the inversion layer	C/cm^2
Q_{ox}	Total charge per unit area in oxide	C/cm^2

Q_{si}	Total charge per unit area in silicon	C/cm^2
Q_W	Synaptic weight charge per unit area	C/cm^2
R	Resistance	Ω
σ	Conductivity	$(\Omega\text{-cm})^{-1}$
t	Time	s
t_{ox}	Oxide thickness	cm
T	Absolute temperature	K
τ	Time constant	s
τ_g	Generation lifetime	s
τ_r	Charging time at membrane node	s
τ_{relax}	Dielectric relaxation time	s
τ_{SID}	Transit time due to self-induced drift	s
τ_{TD}	Transit time due to thermal diffusion	s
υ	Velocity	cm/s
υ_t	Thermal velocity	cm/s
V	Voltage	V
V_{bi}	Built-in potential	V
V_{dd}	Power-supply voltage	V
V_{ds}	Drain-source voltage	V
V_g	Gate voltage	V
V_{gs}	Gate-source voltage	V
V_i	Pre-synaptic spike	V
V_{ji}	Synaptic weight voltage	V
V_o	Output voltage	V
V_{ox}	Potential drop across oxide	V
V_p	Program voltage	V
V_{sat}	Saturation voltage	V
V_t	Thermal voltage (=0.026 V @ 300 K)	V
V_{E3}	Voltage on synapse output terminal	V
V_F	Floating gate potential	V
V_{FB}	Flatband voltage	V
V_{FN}	Coupled voltage of floating diffusion node	V
V_{PSP}	Membrane potential	V

V_T	Threshold voltage	V
V_{Th}	Switching threshold of CMOS inverter	V
W	Channel width	cm
W_d	Depletion layer width	cm
W_{dm}	Maximum depletion layer width	cm
W_f	Depletion layer width at equilibrium	cm
W_{inv}	Effective inversion layer thickness	cm
χ	Electron affinity	eV

CHAPTER 1 INTRODUCTION

Section 1.1 Towards Intelligent Computation System

The creation of the electronic computers and deoxyribonucleic acid (DNA) model of heredity over 50 years ago facilitates the progress in understanding biological intelligence and creating machine intelligence. Since the first computer was created, the machines have gone through six generations [1, 2, 6, 69]. The first four computer generations were mainly built up respectively from vacuum tubes, transistors, integrated circuitry (IC), and Very Large Scale Integration (VLSI). Giant 'brains' that filled air-conditioned rooms were shrunk into briefcases, with the speed of computation (the computer performance) doubled every two years. The fifth generation computer systems are defined as knowledge-information processing systems [1, 2], which signify the presence of artificial intelligence software and entire microprocessors that run in parallel. Today's most powerful supercomputers are all massively parallel systems such as Blue Gene [3], successfully applied to the simulation of half a mouse brain for ten seconds [4] and the neocortex of human brain [5]. The sixth generation is formed by dedicated neural hardware combining neuroscience, cognitive science, computer science and electronics. The machine of the future can be seen as a cooperative computation system that integrates different subsystems, each quite specialized in structure and some supplied with sensors and motors [6, 69].

In 1965, Intel co-founder Gordon Moore saw the future [7]. His prediction, popularly known as Moore's law, states that the number of transistors on a chip doubles about every two years [8]. In the 1970s, the processors such as 8008 and 8086 had no more than 30 thousand transistors, while the Pentium III processor developed in the late 1990s has the number of transistors up to 28 million. In the recent dual-core Itanium processor, there are more than 1 billion transistors. Moore's law has been the guiding principal for the semiconductor industry over the last 40 years. It has fueled the worldwide technology revolution towards greater performance, energy efficiency, and technologies that create new computing solutions. The scaling has reduced the feature

size of transistor below the 50nm regime in today's CMOS technology and has led to an unprecedented level of integration in achieving ever higher performance and low power consumption. In terms of numerical computations, the capabilities of digital computers far exceed those of animal brains. Nonetheless, the most powerful computing systems still cannot approach impressive feats that can be accomplished by nervous systems of animals [9]. In addition, such an exponential progress is not usually sustainable. The rate of change predicts that in around 150 years there will be more memory cells on a square centimeter of silicon than there are atoms in the universe. Gordon Moore indicated [10] that "*It can't continue forever. The nature of exponentials is that you push them out and eventually disaster happens*". As CMOS technology starts to run into its fundamental limit, conventional scaling will no longer be sufficient to continue device performance by creating smaller transistors. Therefore many efforts have been delivered to extend Moore's law such as the innovations in nano-scale [11, 12]. One of the emerging solutions is to explore bio-inspired architectures. International Technology Roadmap for Semiconductors'2005 (ITRS) [13] states that "*mimicking biological behavior in non-CMOS hardware, in particular copying their enormous fan-in/fan-out, is still in its infancy. ...The feasibility of using nano-scale electronic devices and interconnects to implement such massively parallel, adaptive, self-organizing computational models is starting to become an active research area*".

Biological research has accumulated a larger amount of knowledge of the structure and functions within the nervous systems of living organisms. It has spanned the entire range of neural cognition from quantum fields to social interactions combined with the neural dynamics operating across all levels. A century has passed since the first introduction of the word "synapse" which defines a functional junction between nerve cells. The biological cell is a goal-oriented organism that has the features of adaptability, interactivity, robustness, flexibility, and autonomy [13]. A typical neuron possesses 10,000 synaptic inputs on average, and a number of outputs. It can store memory in the pattern and strength of the analog synapses that connect it to other neurons. Inspired by recent advance in neurophysiological methods, mathematical theory, and computing technologies, the artificial neural networks (ANNs) have been developed based on what is known about how human and animal brains acquire and manipulate knowledge about their environments [14]. The bio-inspired computational

systems are built up from simplified representations of biological neurons, interconnected in a complex web-like structure with adaptive synapses. The bio-inspired systems differ from conventional computing systems since they acquire knowledge by a gradual learning process, and represent it as a pattern of weights. Bio-inspired computational systems made by humans can be leveraged to realize a variety of biological functions ranging from sensing to manipulation aimed at recapitulating in vivo biological dynamics and structures. They demonstrate robust processing capabilities in challenging real life scenarios, and are able to offer solutions to mathematically intractable problems, which conventional programmed systems find very hard to deal with, through their ability to be trained for specific tasks [15]. This is one of the reasons for the fast-growing interest in implementing brain-like intelligence as large neural systems using either hardware or software.

Since neural systems have many nonlinear elements with massively parallel interconnections in nature, building intelligent systems requires large computational resources in terms of speed, power and area. The realization of such networks in software is by far the most common manifestation of this computational technique. However, there are many applications where neural computations running on software platforms may not be an option, for example biological implants [16-19], autonomous robots [20, 21], sensors [22-24], etc., and a dedicated hardware implementation route would then be the preferred option to provide a self-contained and physically robust solution. More importantly, to preserve the parallel processing capability of neural systems and consequently expand the application domain to real time processing, it is necessary to develop efficient hardware implementation techniques. The number and variety of novel neural network paradigms are increasing rapidly though capabilities of many brain-like paradigms are hardly known yet. The more is learnt about the neural systems, the more it is clear that the computations being performed and the hardware being used are quite different from the conventional computers. If very large scale, highly parallel hardware implementations of artificial neurons become a reality, it is essential that neurons with small physical dimensions are made available to facilitate this. The best efforts at biological scale emulation, while functional, are still basic and consume large amounts of power. Therefore, there is a pressing need for small geometry, low power neural building blocks with operational characteristics that closely mimic realistic neural systems. Continuing research on hardware

implementation of neural networks by efficiently utilizing the similarities between silicon and biophysics, can create new opportunities in advancing the understanding of the biological mechanisms and exploring novel ideas in massively parallel computational systems.

Section 1.2 Neural Networks

In this section, the fundamentals of biological neural networks are presented. A brief review of history of artificial neural networks (ANNs) is then given. The newly developed spiking neural networks (SNNs) and its neuron models which take the advantages over the conventional techniques are described in detail. The operation rules of the synaptic plasticity especially the spike-timing dependent plasticity (STDP) are also presented in this section.

Section 1.2.1 Biological Neural Networks

In human cerebral cortex, there are about of the order of 10^9 neurons properly interconnected by synapses of the order of 6×10^{12} in a network. Neurons are the elementary processing units in the brain and are typically eight to nine orders of magnitude slower than silicon logic gates. However, as a result of the massive interconnections and parallelism, the brain compensates for the relatively slow operation of a neuron [14]. Inter-neuron communication takes place via action potentials induced by the voltage-dependent currents that pass through ion channels in the cell membrane. A synapse is an electrochemical contact between biological neurons, each of which has three main components, called cell body (or soma), axon, and dendrites as depicted in Fig. 1.1.

The soma of the neuron receives the sum of the input currents which arrive at the dendrites. Once the membrane potential at the cell body exceeds a certain threshold, the neuron generates a millisecond-long pulse, called an action potential or spike. This

spike is in turn conveyed through the axon, which ends in a number of synapses, to the dendrite trees of other neurons. The most common type of synapse is the chemical synapse, which functions by converting a pre-synaptic electrical signal into a chemical signal and back into a post-synaptic electrical signal. In a chemical synapse, the axon of the pre-synaptic neuron is extremely close to the dendrite of the post-synaptic neuron.

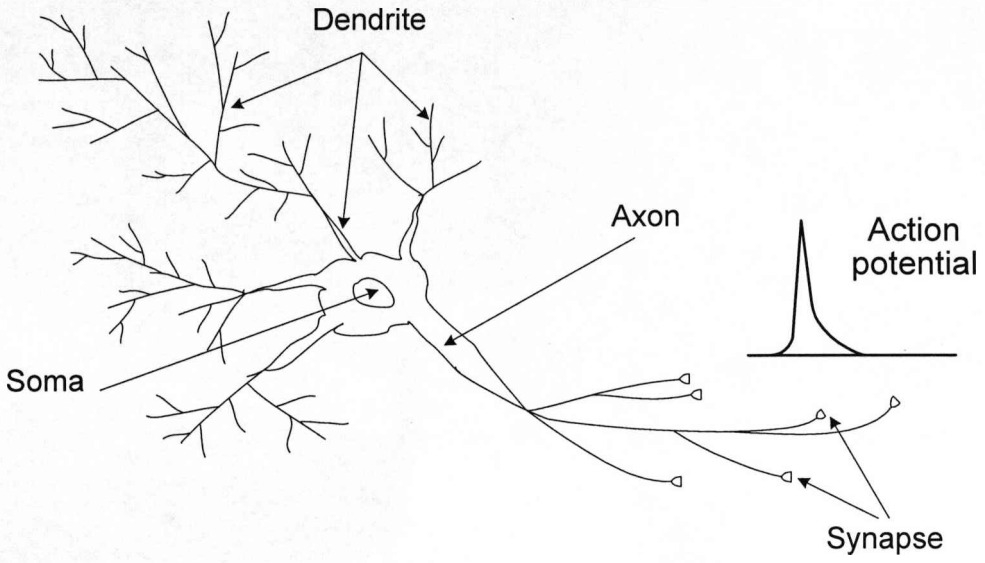


Fig. 1.1 A biological neuron, consisting of a soma, many dendritic trees, and an axon with many axon terminals. A synapse is the junction between an axonal terminal of a pre-synaptic neuron and a dendrite of a receiving neuron.

Neural networks learn by changing the strength of the synapses that can be differentiated into two categories: Excitatory, where the neurotransmitter increases the membrane potential of a neuron; Inhibitory, where the neurotransmitter decreases this potential. At rest, the neuron membrane typically has a potential of -65mV. Increasing the potential is referred to depolarising since it reduces the negative potential; conversely decreasing the potential to the resting potential is referred to repolarising. Hyperpolarising is referred to the situation when the membrane potential drops below the resting potential, effectively providing a refractory period. Fig. 1.2 shows the summation of post-synaptic potentials (PSPs) consequently producing a spike when

the threshold is exceeded. Note that $t_1^{(1)}$ and $t_1^{(2)}$ are the first and second PSPs respectively, generated by neuron 1, and $t_2^{(1)}$ and $t_2^{(2)}$ are the PSPs generated by neuron 2 [25, 28, 29].

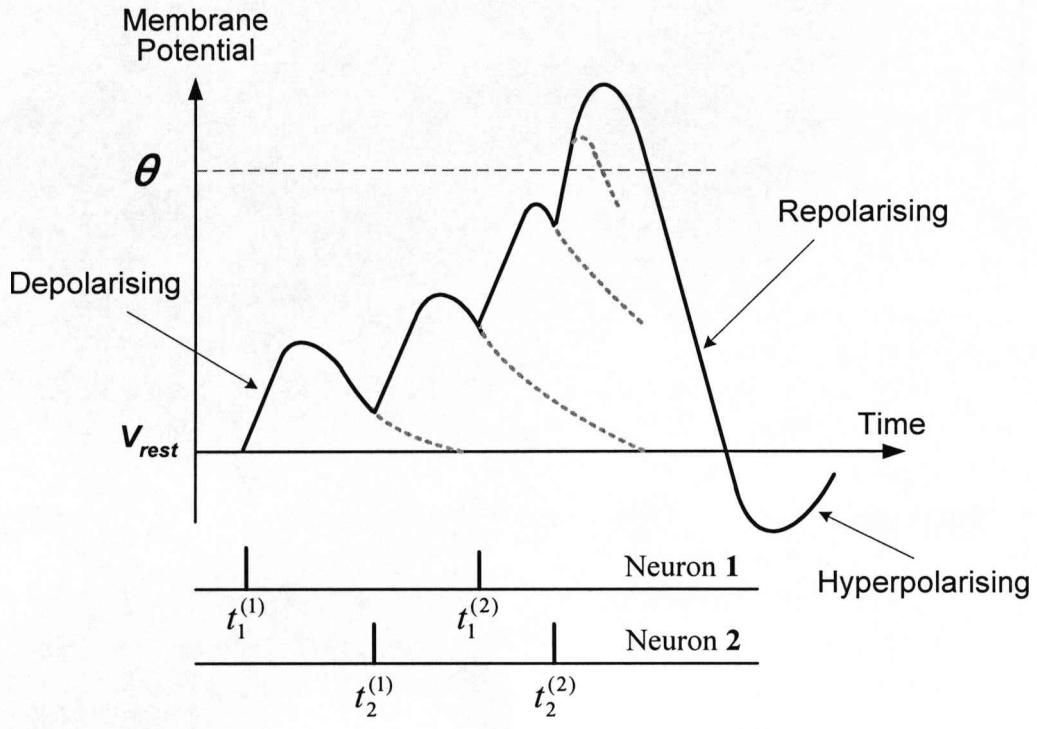


Fig. 1.2 Spike productions from PSP summation.

Section 1.2.2 Artificial Neural Networks

ANNs, introduced to model functions and performance of the central nervous systems in living creatures, represent a promising technology that is rooted in many disciplines: neurosciences, mathematics, statistics, physics, computer science and engineering [14]. In the most general form, an artificial neural network is designed to replicate the way in which the biological nervous systems perform a particular task or function. It can acquire, store, and use experiential knowledge to learn from and adapt to the environment in real time. ANNs are usually comprised of a large number of simple, interconnected processing units that are analogous to biological neurons which work in massive parallelism. They are connected together with weighted

junctions that are analogous to biological synapses. The neurons process and transform the input signals that can be transmitted to the others. The encoding of information in the network is achieved during the learning phase (supervised, unsupervised or reinforced), where the synaptic weights existing between the neurons are modified in response to external stimuli [14].

Since the introduction of the first ideas and models, ANNs have gone through three generations, distinguished according to their computational units [27]. The first generation of ANNs focuses on McCulloch-Pitts threshold neurons which can only produce digital outputs. These neurons encode information by the presence or absence, rather than the shape of the spikes, and successfully give rise to multi-layer perceptrons and Hopfield nets. The second generation artificial neurons use the continuous activation function and can process information with analog inputs and outputs. The neural models are able to integrate training algorithms that are based on gradient descent. Typical examples of such ANNs are feedforward and recurrent sigmoid networks, and radial basis function networks [27]. The outputs of the neuron models of the first two generations are viewed as the mean firing rates of biological neurons. From a biological point of view, it has conventionally been thought that information was contained in these mean firing rates of the neurons.

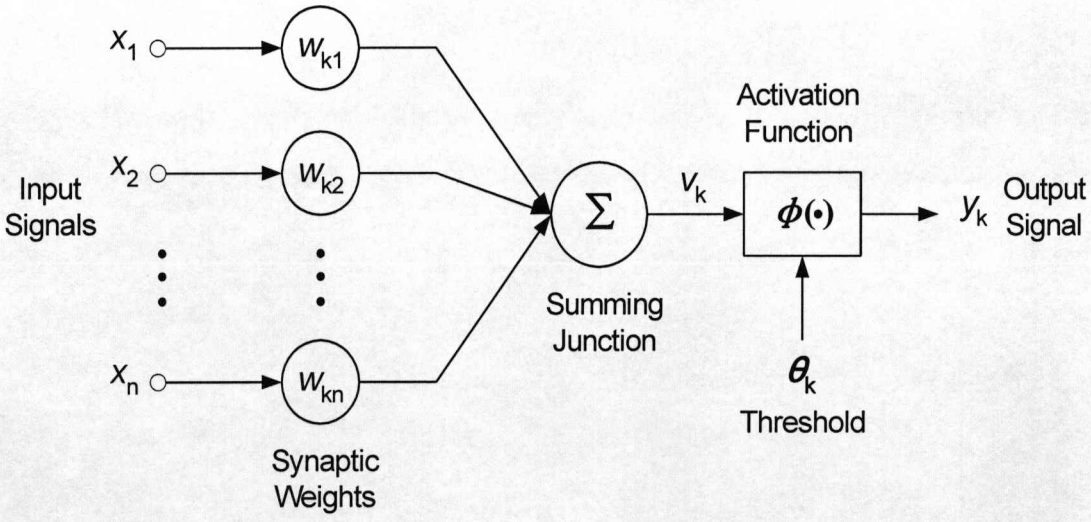


Fig. 1.3 Non-linear model of a neuron. $X_1, X_2, \ldots, X_n$ are the input signals; $W_{k1}, W_{k2}, \ldots, W_{kn}$ are the synaptic weights of neuron k; θ_k is the threshold; $\phi(\bullet)$ is the activation function; Y_k is the output signal of neuron k.

An artificial neural network can be described with a mathematical model shown in Fig. 1.3. In this model, synapses are seen as multipliers that modulate the amplitude of the pre-synaptic spikes by the weight values of the synapses. The accumulative effect of the membrane potential is thought as a summation of the multiplied spikes in time. An activation function (step, sigmoid or spike) limits the output of the neuron to a finite value. The activation function may also provide the model with non-linearity [14].

Section 1.2.3 Spiking Neural Networks

Recently experimental evidence has accumulated to suggest that the mean firing rate method could not describe brain activity. The main reason is that the reaction times in behavioral experiments are too short to allow temporal averaging for the mean firing rate. Instead it has been suggested that the timing of single spikes is used to encode information [26-29], and the spike-based coding schemes are considerably more efficient than the rate code methods [30]. The relative timing of the spikes could also encode information in addition to the single spike independently. Synchronized firing of neurons allows the brain to perform pattern segmentation, feature binding, and figure separation [31, 32]. This has led to the third generation of neural networks, so-called spiking neural networks, which is based on spiking neurons as the computational elements. A spiking neuron computes by transforming dynamic input into a train of spikes rather than rate codes, so allowing the incorporation of spatio-temporal information in computation and communications. Networks of spiking neurons are proven to be more biologically plausible and computationally powerful than conventional formalisms of ANNs [27].

Spiking neurons describe the inputs in terms of single spikes each of which has amplitude of the order of 100mV and a typical duration of 1 or 2ms as shown in Fig. 1.4. Incoming spikes are weighted and a post-synaptic potential is induced according to an impulse response function, which may be composed of a sharp rise and a slowly decaying function [28]. The form of the spike is invariant as it propagates along the axon. Therefore it is believed that the timing of the spikes rather than spikes

themselves carries the information. A chain of spikes is normally known as a spike train where subsequent to each spike there is a period of time, a refractory period. Within this period, another spike cannot be produced, regardless of the input.

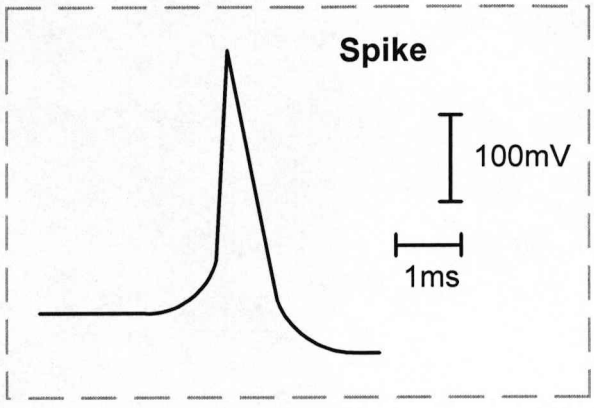


Fig. 1.4 Form of a spike, which has amplitude of the order of 100mV and a typical duration of 1 or 2ms.

The transmission of a single spike from one neuron to another is mediated by the synapse where the two neurons interact. When the pre-synaptic spike arrives at the synapse, the neurotransmitter which influences the membrane potential of the post-synaptic neuron is released. If the membrane potential goes beyond a threshold, the post-synaptic neuron generates an output spike and then the membrane potential is reset by a refractory response. The degree of the influence on a pre-synaptic spike is determined by the type and weight of the synapse.

Section 1.2.4 Spiking Neuron Models

By performing extensive experimental studies on the giant axon of the squid, Hodgkin and Huxley [33] succeeded in measuring the ionic currents that can result in action potentials in the cell membrane, and illuminating their behavior by using four differential equations. In their model, neural activities are described at a microscopic level where the ion channels and conductance along with the membrane potential are

considered. The Hodgkin-Huxley model reproduces not only the exact shape of spikes but also other physiological properties such as refractoriness and repetitive firing. It is a mathematical framework for modeling neural excitability, and an electronic circuit is designed to represent these ionic currents [28, 29]. However, Hodgkin-Huxley model does not easily clarify the intrinsic behaviors underlying the dynamic properties of the model and it is not as popular as some of the other models of experimental research due to its complicated nonlinearity and computational expense. The simplified two-dimensional neuron models, which are mathematically tractable, would provide an intrinsic understanding of the dynamical structures of neural excitability and network behaviors [34, 35].

A phenomenological model at a higher level of abstraction, the so-called leaky integrate-and-fire (I&F), considers the neurons as a homogeneous unit which generates spikes if the total excitation is sufficiently large. It is a simplification of the Hodgkin-Huxley model. The leaky I&F model performs the summation of the multiplied spikes in time, and will fire a spike when the potential across the membrane rises above a threshold value [28]. The cell model basically consists of a capacitor in parallel with a resistor, and a threshold device, as shown in Fig. 1.5.

In response to a stimulant current $I(t)$, the capacitor is charged. The resistor provides a pathway for the leakage current which drives the membrane potential to the resting potential in the absence of an input current.

The stimulus current $I(t)$ can be expressed as:

$$I(t) = \frac{v(t)}{R} + C\frac{dv}{dt} \tag{1.1}$$

where $v(t)$ is the membrane potential across the capacitor. Introducing the time constant τ of the leaky integrator yields:

$$\tau\frac{dv}{dt} = -v(t) + RI(t) \tag{1.2}$$

If the membrane is stimulated by a constant current $I(t) = I_0$ and the resting potential is at zero, then the membrane potential is:

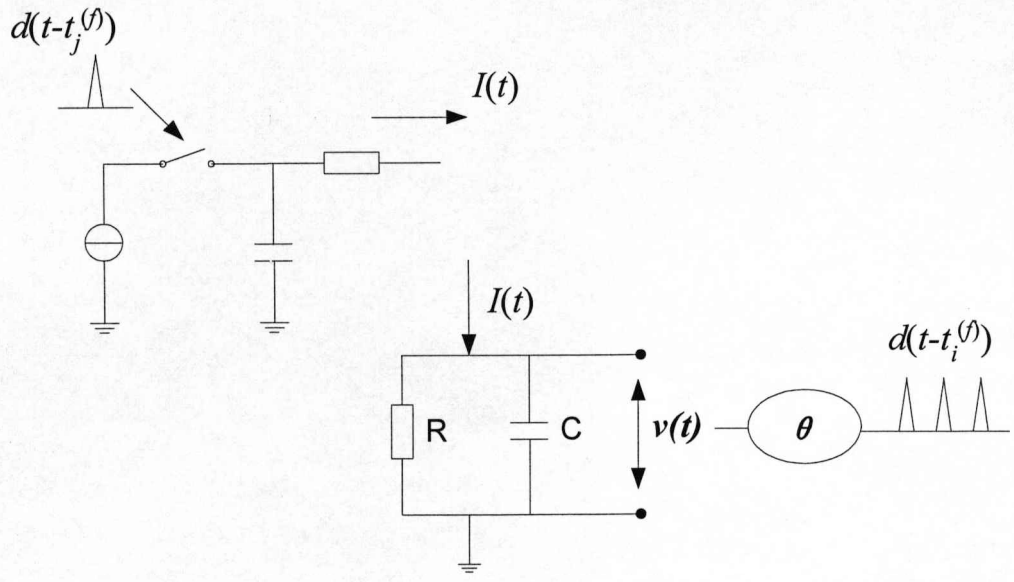


Fig. 1.5 Schematic drawing of the leaky integrate-and-fire (I&F) neuron model, which consists of a resistor *R* in parallel with a capacitor *C* driven by *I(t)*. The resistor provides a pathway for the leakage current to flow. A threshold device is employed to produce a spike when the membrane potential surpasses the threshold, and subsequently the membrane potential is reset to a value below the threshold [28].

$$v(t) = RI_0\left[1-\exp\left(-\frac{t-t^{(1)}}{\tau}\right)\right] \tag{1.3}$$

where $t^{(1)}$ is the time of the first spike occurrence.

For a value of $v(t)$ less than the threshold, θ, no spike can be produced. When $v(t)$ exceeds θ, a spike can occur at time $t^{(2)}$, which can be found by considering the threshold condition:

$$\theta = RI_o\left[1-\exp\left(-\frac{t^{(2)}-t^{(1)}}{\tau}\right)\right] \tag{1.4}$$

The inter-spike interval (ISI), $T = t^{(2)} - t^{(1)}$, can be found by solving (1.4):

$$T = \tau \ln\frac{RI_o}{RI_o-\theta} \tag{1.5}$$

After a spike is produced at $t^{(2)}$, the membrane potential returns to the resting potential and the integration process starts again. The leaky I&F neuron fires regularly with period T for a constant stimulus current [29].

Section 1.2.5 Synaptic Plasticity

Synaptic plasticity is a form of local learning of neural networks by updating the synaptic efficacy either in long term or short term [28, 30, 36]. It is based on Hebbian's postulate which states that where one cell's firing repeatedly contributes to the firing of another cell, the magnitude of this contribution will tend to increase gradually with time. Synaptic plasticity depends on relative timing of spikes. It has been suggested that changing the relative timing of the pre-synaptic and post-synaptic spikes can determine whether a synapse is potentiated or depressed [37-39]. The synaptic weights between learning neurons are adjusted so that each weight can better represent the correlation of firing activity between pre-synaptic and post-synaptic neurons. However pure synaptic plasticity is difficult to control, since it endlessly strengthens the effective synapses and weakens the ineffective synapses by a positive feedback process, resulting in the unstable post-synaptic firing rates. Thus a biologically plausible local rule is required to avoid such behavior [40].

Spike-timing dependent plasticity (STDP) is a form of competitive Hebbian learning where the post-synaptic neurons are sensitive to the timing of the incoming spikes, resulting in competition among the pre-synaptic neurons [36, 40-43]. The relative timing of the pre-synaptic and post-synaptic spiking determines the direction and the degree by which the synaptic efficacy changes. As illustrated in Fig. 1.6, the choice to strengthen or depress the synaptic weight depends entirely on which neuron fires first. Long-term potentiation (LTP) occurs if pre-synaptic action potentials arrive before the post-synaptic spike and long-term depression (LTD) occurs if post-synaptic firing occurs prior to pre-synaptic firing. This learning mechanism has the advantages of shorter latencies, spike synchronization and faster information propagation through the entire neural networks [40]. It can be used to learn temporal delays with high precision and has been used to model visual processing in vision systems [44].

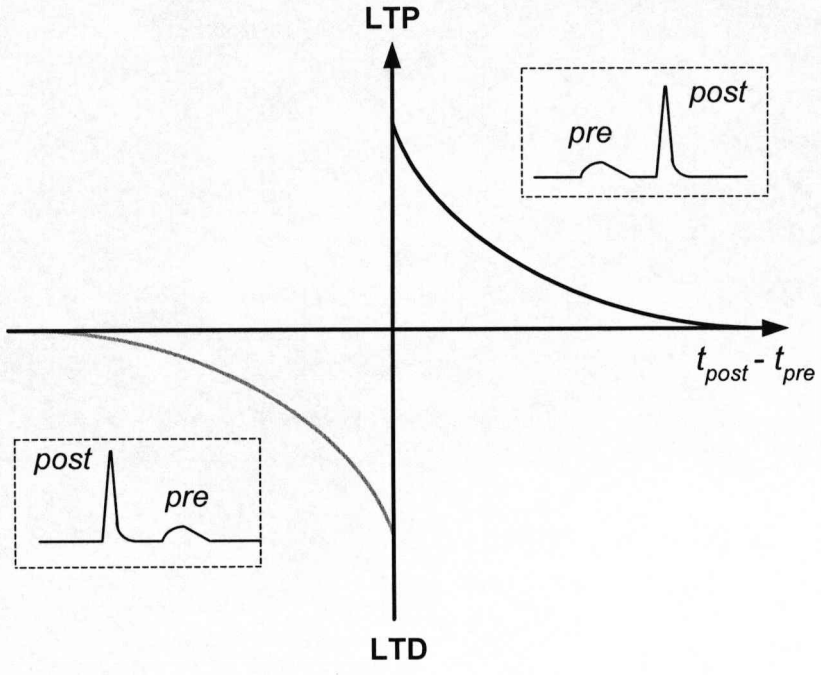


Fig. 1.6 Schematic of STDP function. Maximal plasticity is produced by short negative or positive ISIs.

Section 1.3 Neural Networks in VLSI

Recent development on VLSI systems spans a broad range including semiconductor devices, ICs, digital logic, full VLSI systems and its applications. It provides a powerful resource for implementing neural systems, and gives rise to a multidisciplinary engineering field known as neuromorphic electronics [25, 45-62]. This technology, mimicked from the physical properties and information processing characteristics of the neural systems, aims to reproduce biological computation by translating its anatomy and physiology into custom silicon chips.

Semiconductor devices possess some similar attributes that can facilitate the modeling of neural systems allowing the exploitation of the parallelism associated with these systems. In both semiconductor electronics and nervous tissue, information is manipulated on the basis of charge conservation. Electrons in semiconductor and ions in nervous tissue are both in thermal equilibrium with their surroundings, and their

energies are Boltzmann distributed [25, 63]. The fundamental forces causing ion flow in biology are the same fundamental forces causing electron flow in a MOS transistor operating with low currents [25, 64]. The nervous system has mechanisms for long-term memory and for learning, including synaptic plasticity and neuronal growth [65]. Semiconductor electronics also has mechanisms for long-term memory.

Today a wide range of implementation techniques can be chosen for neuromorphic hardware: analog, digital, or hybrid design. Special attention is given to the adaptability, flexibility, scalability, and great increase in speed over conventional computing systems. Analog technique can exploit physical properties to perform neural functions and thereby obtain high speed and densities. Since the analog versions of the neural operation such as multiplication can be much smaller than their digital equivalents, massively parallel neural systems can be efficiently implemented by using analog VLSI, giving a potential for real time operation. The use of MOS transistors in subthreshold regime enables the systems to operate at lower power. Although digital systems can function in subthreshold, analog systems carry more information per wire and fewer transistors per operation and thus inherently dissipate less power [66]. Furthermore, analog VLSI directly interfaces with the real world, eliminating the need for fast analog-to-digital and digital-to-analog converters [25]. Digital techniques have the advantages of high programmability, high computational precision, and high reliability [69]. A number of powerful design tools and kits are readily used for digital full- and semi-custom design [67, 68]. Moreover, synaptic weights can be stored and updated on or off chip, determined by the trade-off between speed and size [69]. The main disadvantages of digital techniques are the relative low computational speed and the large power and area consumption of neural circuits. Disadvantages of analog techniques are the sensitivity to noise and interference and susceptibility to process variations. The hybrid of the analog and digital techniques will be advantageous and suitable for the hardware implementation of neural networks [52, 58-60, 62, 69-71]. Using hybrid VLSI circuits enables neuromorphic engineers to build dense integrated networks of silicon neurons that run in real time, and to capture the computational power and efficiency of biological neural systems.

Early experiments have shown that the elementary operations of the nervous system can be emulated by analog circuits for the creation of novel devices. Thereafter a

number of silicon models of specific areas of the nervous system have been built to perform massively parallel signal processing, such as the retina chip [47, 56, 72-76], which is modeled after human vision, as well as silicon cochlear [77, 78], auditory midbrain [79], motion sensor [23, 80-83], and olfaction chip [84, 85]. The silicon retina by Mahowald and Mead [47] detects the contours of a moving stimulus, and generates only analog output values. The silicon retina by Zaghloul and Boahen [56, 76], contains a 60×96 array of phototransistors and processing circuits and is able to generate spiking outputs that mimic the responses of ON-sustained and OFF-sustained retinal ganglion cells. This retina chip is suitable for multichip neuromorphic systems. More recently, Koickal et al. [85] presented the analog VLSI implementation of the circuit building blocks of an adaptive neuromorphic olfaction chip, with on-chip chemosensor array and on-chip sensor interface. Their method integrates an on-chip spike time dependent learning circuit to dynamically adapt weights for odor detection and classification. Rasche [61] described an adjustable and excitable membrane, consisting of spiking units. It is designed to fit into a multichip neuromorphic system, and can perform different visual tasks such as contour detection, contour propagation, image segmentation, and motion detection [86].

To implement higher levels of processing and cognition such as the models of cortical regions, a number of multichip approaches and proper communication protocol between chips have been reported [59, 62, 87-90]. These neuromorphic systems use a similar design strategy as biological systems: local computations are performed in analog, and the results are communicated by using all-or-none spikes [62]. The common language of neuromorphic chips is the address–event representation (AER) communication protocol [60, 91-97], originally introduced by Mahowald and Sivilotti. The protocol employs time-multiplexing to mimic extensive connectivity between biological neurons. An address encoder generates a unique binary address for each neuron whenever it spikes. A digital bus transmits these addresses to the receiving chip, in which an address decoder selects the corresponding location [60]. The protocol is asynchronous, with the time that the address appears on the bus encoding the spike time directly. Choi et al. [89] proposed a neuromorphic multichip implementation of orientation hypercolumns in the mammalian primary visual cortex, which consists of a single silicon retina feeding multiple orientation selective image

filtering chips [88]. Each chip contains a 2-D array of neurons tuned to the same orientation and spatial frequency, but different retinal locations. All chips operate in continuous time, and communicate with each other using spike-encoded inputs and outputs which are transmitted by the digital asynchronous AER protocol.

Section 1.4 Silicon Neuron

Biological neurons are the basic processing units of the neural networks. Its computational behavior can be explained by various spiking neuron models, rather than the traditional rate-based models [35]. The conductance-based models, such as the Hodgkin-Huxley model, are usually used to capture the detailed biophysical dynamics of the neurons. Whereas simpler models such as the integrate-and-fire (I&F) model, are sufficient for the case where the realistic and detailed spiking mechanisms are not required, and this type of models are widely used for large scale network simulations. Many silicon circuits that emulate the spiking neuron have been reported in the literature, and analog spiking neuron circuits have been commonly used for mimicking low-level neural operation such as sensory occurring in cochlea, retina, and the visual cortex. Conventionally silicon neurons have been designed according to two major principles: the conductance-based design [48, 98-102], faithfully mimicking the nonlinear channel conductance in a real neuron; and the phenomenological design [25, 55, 58, 103-105], implementing very simplified neuron models.

The former principle is employed to reproduce the electrophysiological properties of biological neurons more accurately. The conductance-based neuron circuits can achieve the complex behavior that is experimentally observed in living neurons. However, this kind of circuits is not suitable for a model based approach since the intrinsic characteristics of neurons cannot be easily deduced from a complex network with many neural parameters. Moreover, the circuits tend to take up a lot of silicon area if the theoretical models to be implemented are incompatible with the characteristics of electronic devices, failing to reach the number of spiking neurons necessary to be incorporated in system level phenomena. In [48], Mahowald and

Douglas proposed sophisticated neuron circuits which are analogous to the Hodgkin-Huxley model. This approach can yield highly detailed single neuron models, but seems to have used many approximations to the underlying dynamics. Dupeyron et al. developed multiple integrated circuits to implement a silicon neuron that matches well the dynamics of living neurons [98], but inefficient in both size and power. The work led to many improved implementations, incorporating both the spike mechanism and the adaptation mechanism, such as the implementation of a cortical pyramidal cell with about 30 adjustable parameters on a $4mm^2$ chip in a standard 1.2μm CMOS technology [100], and the six-conductance implementation with 5 silicon neuron on a $4mm^2$ chip using the same process [101]. A recent publication [102] describes a network of silicon inter-neurons that synchronize in the gamma frequency range (20-80Hz), which strongly influences neuronal spike timing within many brain regions, potentially playing a crucial role in computation. Each inter-neuron, including shunting inhibition (conductance-based) and the synaptic delay, requires 22 transistors. By using the similarities between biological and silicon channels, Farquhar and Hasler developed a neuron circuit which accurately models action potentials and channel currents of real neurons [64]. The circuit requires four capacitors and six transistors, which is significantly smaller than systems emulating model equations [48], while also closely modeling biological physics.

Although theoretically it might be possible to incorporate all realistic details of neurons into the silicon model, one shall have to consider a tradeoff between the complexity of individual silicon neuron and the number of neurons on a single die. In most cases, building a single neuron is only a step to simulate the interactive behavior of a large network of neurons. Therefore it is necessary to minimize the amount of detail and reduce the circuit size, in order to allow more neurons to be put on a single die.

The phenomenological design is employed to reproduce certain properties of neurons that are considered to be especially important by designers. Typical examples of such design are I&F silicon neurons, where the spikes are generated whenever the membrane potential reaches a threshold. This kind of neuron circuits excludes some intrinsic dynamic behavior in the operation of real neurons, such as the refractoriness and variable delay. Although the simplified models ignore the detailed time course of

action potentials, they are widely used for studies on neural coding, and large network dynamics. The simple axon-hillock circuit by Mead consists of an integrating capacitor connected to two inverters, a feedback capacitor, and a reset transistor driven by the output inverter [25]. The membrane potential is implemented as the voltage across the integrating capacitor. A spike is generated when the integrated voltage exceeds the switching threshold of the first inverter. This circuit is widely used in the VLSI realization of neural networks [55]. However, the circuit dissipates significant amounts of power since the first inverter spends a large amount of time in the region in which both transistors conduct a short-circuit current. A further drawback is that the switching threshold of the inverter depends only on process parameters, and does not model additional neural characteristics, such as spike frequency adaptation or refractory period mechanisms. While Schultz and Jabri proposed an alternative approach to implement spike frequency adaptation [103], their neuron circuit with an explicit threshold voltage has large power consumption for the same reasons as the axon-hillock circuit. A low power neuron circuit without adaptation is reported by van Schaik, using an amplifier at the input that compares the voltage on the capacitor (membrane potential) with a desired spiking threshold voltage [104]. When the input exceeds the threshold, the amplifier drives the inverter strongly, making it switch very rapidly. Recently more low power adaptive I&F neuron circuits have been reported in the literature [58, 105]. These circuits are very compact, using approximately 20 transistors, but further reduction is required towards biological-scale neural networks in silicon.

Section 1.5 Silicon Synapse

Synapses are connecting junctions between neurons that play an important role in development, memory and learning of neural networks [40]. Synaptic dynamics is widely believed to be crucial for learning neural codes and encoding spatio-temporal spike patterns. Changes in the synapses contribute to memory storage, and the activity dependent development of neural networks. Within the human nervous system an individual neuron may have 10^4 associated synapses, and consequently the physical space occupied by silicon synapses dominates the neural networks in a single chip.

Therefore it is an important consideration to design a silicon synapse, used in VLSI neural networks, with least area and energy consumption.

Several silicon devices and analog VLSI circuits modeling synaptic functionality have been proposed in the past. In the late eighties, Mead proposed a pulsed current-source synapse, which can be activated by an active-low input spike [25]. The synapse is compact, and has been extensively used in the hardware implementation of spiking neural networks [28, 55, 106]. However this synapse circuit cannot distinguish input spike trains with same mean firing rates but with different timing distributions, and does not produce continuous output signal. While the synaptic circuit described in [107] can produce the decaying output current in response to the input spike, by using three transistors and one capacitor, but its response depends only on the last input spike, failing to reproduce the linear summation of the post-synaptic currents. A modified synaptic circuit proposed by Arthur and Boahen [108] has 4 saturated transistors operating in subthreshold and a capacitor. In this circuit, the dynamics of synaptic output depend on the total number of spikes received, but the linear integration cannot be performed. One of its variants, the log domain integrator synapse has been proposed [109]. It exploits the logarithmic relationship between gate-source voltages and channel currents of transistor in subthreshold. Since the synaptic circuit can implement the linear integrator, the same synapse can perform the linear summation of the spikes arriving from different neurons. The main disadvantage of the circuit is that the spike duration typically lasting less than a few microseconds are too short to inject enough charge in the membrane capacitor of the post-synaptic neuron. Boahen [74] also presented a current mirror integrator synapse, which can generates a mean output current that is saturating nonlinearly with the maximum amplitude and increases with the firing rates. However this synapse circuit cannot be used to aggregate the post-synaptic currents linearly. The response properties of the circuit have been analytically derived for steady state conditions [110], and this synapse has been widely used by some neuromorphic projects [111, 112].

The important characteristic of the biological synapse is the ability of learning by modifying the synaptic weight and its adaptability to an unknown environment. Many circuits has been proposed for mimicking synaptic plasticity, both on short time scales

with models of short-term depression (STD) [113, 114], and on long time scales with spike-based learning schemas, such as STDP [115-118]. Chicca et al. [119] presented a silicon synapse with short-term adaptation consisting of 14 transistors and two capacitors. Two current mirror integrators form the core of the synaptic circuit that integrate the input spikes and produce facilitating/depressing signals with different dynamics. Their silicon synapse, including the interfacing circuitry, is suitable for multichip systems communicated using the AER protocol. However the current mirror integrators are sensitive to device mismatch, though there is a possibility to improve the reliability by increasing the size of transistors. In [57], a feedforward network of silicon neurons with STDP synapses has been developed. The implemented learning rule can be tuned to have a moderate level of weight dependence. This helps stabilize the learning process and still generates binary weight distributions. A more powerful synaptic circuit implementing both long-term and short-term plasticity has been proposed [105]. The weight of any synapse in the array can be changed by setting the pre- and post-synaptic mean firing rates to appropriate values. The STD sub-circuits in the synapses can be activated during or after learning to implement local gain control mechanisms and introduce an additional degree of adaptation.

However, current VLSI architectures fail to match the scale of biological networks because the fundamental building blocks (transistors) require complex circuitry to emulate a synapse. A promising approach based on floating gate technology to implement the synapse in a single device has been developed [120-123]. Floating gate devices are not just the digital memories, but the computational units with analog memory and time-domain dynamic characteristics all in a single device. The single transistor based synapses provide long-term nonvolatile analog storage, multiplication, and local adaptation in silicon: hot-electron injection and electron tunneling permit bidirectional memory updates. In [124], Diorio et al. demonstrated computation and unsupervised learning in an array of floating gate MOS synapse transistors. The learning behavior is described using rules derived directly from the silicon-MOS and silicon-oxide physics. The synaptic array can achieve fast computation and slow adaptation: The inner product computes in 10s, whereas the weight normalization takes minutes to hours. The work presented in [125] investigated the weight dynamics of the floating gate *p*FET synapse, and derived a

weight updating rule such that the equilibrium weight value is proportional to the correlation between the gate and drain voltages. An 11-transistor automaximizing bump circuit has been developed to implement a similarity computation, local adaptation, simultaneous adaptation and computation and nonvolatile storage [126]. This circuit can be used as a core building block for implementing competitive learning networks. Although the single transistor synapse is small and is operated at subthreshold level, allowing the development of dense, low power silicon learning systems, its solid-state characteristics are not compatible with spiking neural networks where the time constants associated with the output transient of the loaded synapse as well as plasticity play vital computational roles within the neuron cell.

Section 1.6 Overview of the Project

The aim of this project is to develop building blocks for the hardware implementation of spiking neural networks. The synapse architectures can be considered to fall into two classes. The first class is based on Charge Coupled Device (CCD) technology which features low doped substrate regions and thick gate oxides. CCD processes allow the formation of relatively long channel transistors which are preferred to build the analogue circuit aspects of the proposed approach. The second class can be fabricated in standard CMOS process which has the great advantage of very low cost, is easily accessible and offers the propensity for usage of the circuit block designs and concepts by other researchers.

The project starts with the idea of a two-phase charge coupled synapse which has the potential to emulate biological plasticity of the spiking neural networks in hardware. Theoretical and preliminary simulation studies show that the proposed silicon synapse exhibits a spike characteristic (a sharp rise and a slower decay) akin to biological synapses, and the degree of the influence on a pre-synaptic spike is determined by the synaptic weight. A full investigation of the proposed synapse is presented which serves to demonstrate that the synapse is able to capture a number of intrinsic properties of the biological synapses. The slow decay of the post-synaptic potential (PSP) in milliseconds is replicated and can be tailored over the range of 0.6ms to 6ms

to provide realistic PSP dependency. Moreover, the overall relaxation of an activated synapse to equilibrium in a few seconds can be thought of as mimicking a refractory period, although the biological equivalent normally lasts of the order of milliseconds. The structure has two further, major drawbacks however. Firstly the PSP characteristic of the synapse depends on the carrier generation lifetime which makes it very dependent on the fabrication process. The decay time constant depends on the carrier generation rate, that is strongly temperature dependent which may not be a major problem given the intended very low operating power levels of the full system. A further major disadvantage is that the structure cannot be made using a commercial CMOS process making it expensive and therefore unlikely to be adopted by other researchers and so is less likely to make an impact on the field. The structure also features only a limited number of biological phenomena, namely PSP and refractory period. Transient analysis of the MOS capacitor is presented, showing that the synapse can only match the pre-synaptic spike train at operating frequencies of the order of 1Hz. Therefore there is a need to develop structures to speed up the synaptic weight generation process, so as to make the synapse operate in response to the spike trains with realistic inter-spike intervals (ISIs) of milliseconds or preferably, much shorter. The architectures are unlikely to attain the parallelism of nature so a trade-off between parallelism and system speed needs to be used – that is, to exploit the natural high speed performance of electronics.

The next structure investigated, incorporates an additional N+ diffusion, which when forward biased, acts as a controlled source of minority carriers to speed up the response time, previously controlled by thermal generation. The idea builds on the study of the previous proposed synapse which indicated the need for smaller time constants and the facility to tune the relaxation of response over a wide range. The device is investigated by theoretical and simulation studies. It is concluded that this approach can speed up and control the synapse response time by means of minority carrier charge injection from the N+ injector diffusion, into the weight storage well. Therefore this programmable synapse can implement dynamic operation associated with biological synapses by setting the ISIs to microseconds and consequently the operating frequencies up to 1MHz.

To avoid the need for custom design, a further structure is proposed, namely the programmable charge coupled synapse. This structure is also designed to incorporate facilitation and depression features of a biological synapse. The refresh of the synaptic weight charge (potentiation of synapse) is realized by charge injected from a MOS transistor operating in subthreshold. By varying the gate voltage of this transistor, the refresh time of the synapse is tuneable, enabling the programmability of the synapse. This synapse design is compatible with commercial CMOS processing. Transistor parameters from the AMS 0.35μm CMOS mixed-signal process are used in the study. The correlation between the amplitude of the output spike and the level of charge in the storage well, which is used to implement synaptic plasticity, is demonstrated. Most importantly, the synapse is able to facilitate over the frequency range 1-100MHz.

In addition to the development of synapse architectures using semiconductor devices, a neuron cell circuit is proposed and investigated using both theory and PSpice circuit simulation. The slow discharge by reverse-biased diode leakage is expected to mimic the decay of the membrane potential of biological neurons. For the case of a diode with leakage limited by thermal generation, it is feasible to set the decay time constants by lifetime quenching. However, the use of a leaky diode, where the leakage is limited by Zener tunneling, is preferred as it has very weak dependence on temperature. Such zener diodes can easily be formed by using the contacts of the source/drain of *p*- and *n*-MOS transistors, and limit the voltage levels to be less than its 'breakdown' voltage. In fact, the use of diodes results in rather long membrane potential (order of ms) which mimics nature but do not take advantage of the inherently high speed of electronics. Therefore, an alternative approach employing a transistor biased into subthreshold is considered to form a leakage path, making the decay time constant tunable. Parameters of the AMS 0.35μm CMOS mixed-signal process are used to establish Spice 'breakout' devices. The output current spikes of the synapses are modeled by a piecewise linear current source in Spice. Electronic emulation of PSP summation and membrane potential depolarizing/repolarising phenomena, as well as the firing behavior, of the biological neuron are observed in the simulation, and the effectiveness of the neuron circuit to incorporate an array of programmable synapses is verified.

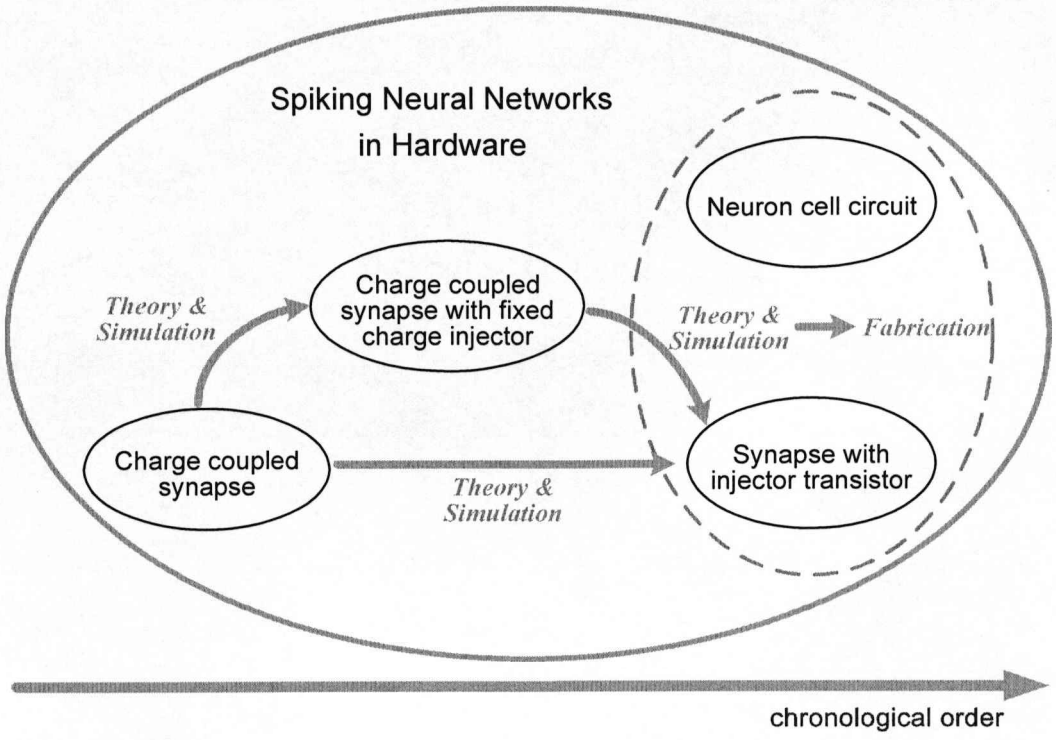


Fig. 1.7 Development of the solid-state building blocks for spiking neural networks in chronological order.

Section 1.7 Organization of the Thesis

The rest of the thesis is organized as follows. Chapter 2 describes the fundamental physics of the basic semiconductor devices employed in this project. In Chapter 3, a charge coupled synapse is proposed, with identified characteristics of biological synapses. The principles and operation are presented and analyzed with the aid of simulation study. Chapter 4 presents the implementation of programmable dynamic synapses in single semiconductor devices. Simulation results are presented which clearly demonstrate its operation. Chapter 5 develops a low power compact silicon neuron cell that accommodates the proposed silicon synapses. The mathematical analysis and simulation are presented to support this work. A summary of the thesis and suggestions for further work are given in Chapter 6.

References

[1] A. Carling, *Introducing Neural Networks*. Wilmslow, UK: Sigma Press, 1992.

[2] P. C. Treleaven and I. G. Lima, "Japan's fifth generation computer systems," *Computer*, vol. 15, no. 8, pp. 79–88, 1982.

[3] A. Gara, M. A. Blumrich, D. Chen, et. al., "Overview of the Blue Gene/L system architecture," *IBM Journal of Research and Development*, vol. 49, no. 2/3, pp. 195–212, 2005.

[4] J. Frye, R. Ananthanarayanan, and D. S. Modha, "Towards real-time, mouse-scale cortical simulations," presented at the Computational and Systems Neuroscience (CoSyNe) Meeting, Salt Lake City, Utah, Feb. 22–25, 2007.

[5] M. Djurfeldt, M. Lundqvist, C. Johansson, et. al., "Brain-scale simulation of the neocortex on the IBM Blue Gene/L supercomputer," *IBM Journal of Research and Development*, vol. 52, no. 1/2, pp. 31–41, 2008.

[6] M. Arbib, "Schemas and neural networks for sixth generation computing," *Journal of Parallel Distributed Computing*, vol. 6, no. 2, pp. 185–216, 1989.

[7] G. Moore, "Cramming more components onto integrated circuits," *Electronics Magazine*, vol. 38, no. 8, April 19, 1965.

[8] G. Moore, "Progress in digital integrated electronics," *IEDM Tech Digest, IEEE*, pp.11–13, 1975.

[9] C. Mead, "Neuromorphic Engineering: overview and potential," Plenary talk at the International Joint Conference on Neural Networks (IJCNN), Canada, 2005.

[10] M. Dubash, "Moore's law is dead, says Gordon Moore," *Techworld*, April 13, 2005. Available: http://www.techworld.com/opsys/news/index.cfm?NewsID=3477

[11] M. Lundstrom, "Moore's law forever?" *Science*, vol. 299, no. 5604, pp. 210–211, Jan. 2003.

[12] D. K. Ferry, "Nanowires in nanoelectronics," *Science*, vol. 319, no. 5863, pp. 579–580, Feb. 2008.

[13] *The International Technology Roadmap for Semiconductors* (*ITRS*), 2005.

[14] S. Haykin, *Neural Networks*: *A Comprehensive Foundation*, 2nd ed., New Jersey: Prentice Hall, 1999.

[15] R. Kozma, H. Aghazarian, T. Huntsberger, E. Tunstel, and W. J. Freeman, "Computational aspects of cognition and consciousness in intelligent devices," *IEEE Computational Intelligence Magazine*, vol. 2, no. 3, pp. 53–64, 2007.

[16] T. W. Berger, M. Baudry, R. D. Brinton, et. al., "Brain-implantable biomimetic electronics as the next era in neural prosthetics," *Proc. IEEE*, vol. 89, no. 7, pp. 993–1012, 2001.

[17] C. Diorio and J. Mavoori, "Computer electronics meet animal brains," *Computer*, pp. 69–75, Jan. 2003.

[18] P. Fromherz, A. Offenhausser, T. Vetter, and J. Weis, "A neuron-silicon junction: a Retzius cell of the leech on an insulated-gate field-effect transistor," *Science*, vol. 252, no. 5010, pp. 1290–1293, 1991.

[19] R. A. Kaul, N. I. Syed, and P. Fromherz, "Neuron-semiconductor chip with chemical synapse between identified neurons," *Physical Review Letters*, vol. 92, no. 3, 038102(4), 2004.

[20] T. K. Horiuchi, "Seeing in the dark: neuromorphic VLSI modeling of bat echolocation," *IEEE Signal Processing Magazine*, pp. 134–139, September 2005.

[21] S. Still, K. Hepp, and R. Douglas, "Neuromorphic walking gait control," *IEEE Trans. Neural Networks*, vol. 17, no. 2, pp. 496–508, 2006.

[22] G. Indiveri, "Neuromorphic analog VLSI sensor for visual tracking: circuits and application examples," *IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process.*, vol. 46, no. 11, pp. 1337–1347, 1999.

[23] S. C. Liu, "A neuromorphic aVLSI model of global motion processing in the fly," *IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process.*, vol. 47, no. 12, pp. 1458–1467, 2000.

[24] A. A. Stocker, "Analog VLSI focal-plane array with dynamic connections for the estimation of piecewise-smooth optical flow," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 51, no. 5, pp. 963–973, 2004.

[25] C. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.

[26] F. Rieke, D. Warland, R. R. de Ruyter van Steveninck, and W. Bialek, *Spikes – Exploring the Neural Code*, Cambridge, MA: MIT Press, 1997.

[27] W. Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[28] W. Maass and C. M. Bishop, *Pulsed Neural Networks*. Cambridge, MA: MIT Press, 1999.

[29] W. Gerstner and W. M. Kistler, *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002.

[30] S. J. Thorpe, A. Delorme, and R. VanRullen, "Spike-based strategies for rapid processing," *Neural Networks*, vol. 14, no. 6-7, pp. 715–726, 2001.

[31] R. Eckhorn, R. Bauer, W. Jordan, et. al., "Coherent oscillations: a mechanism of feature linking in the visual cortex?" *Biological Cybernetics*, vol. 60, no. 2, pp. 121–130, 1988.

[32] C. Von der Malsburg, "The correlation theory of brain function", in *Models of Neural Networks II*, Domany et. al., Eds., London, UK: Springer-Verlag, 1994, pp. 95–119.

[33] A. L. Hodgkin and A. F. Huxley, "A quantitative description of ion currents and its application to conduction and excitation in nerve membranes," *Journal of Physiology*, vol. 117, pp. 500–544, 1952.

[34] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Trans. Neural Networks*, vol. 14, no. 6, pp. 1569–1572, 2003.

[35] E. M. Izhikevich, "Which model to use for cortical spiking neurons?" *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1063–1070, 2004.

[36] L. F. Abbot and S. B. Nelson, "Synaptic plasticity: taming the beast," *Nature Neuroscience Review*, vol. 3, pp. 1178–1183, 2000.

[37] H. Markram and M. Tsodyks, "Redistribution of synaptic efficacy between neocortical pyramidal neurons," *Nature*, vol. 382, pp. 807–810, 1996.

[38] H. Markram, J. Lubke, M. Frotscher, and B. Sakman, "Physiology and anatomy of synaptic connections between thick tufted pyramidal neurons in the developing rat neocortex," *Journal of Physiology*, vol. 500, pp. 409–440, 1997.

[39] R. S. Zucker and W. G. Regehr, "Short-term synaptic plasticity," *Annu. Rev. Physiol.*, vol. 64, pp. 355–405, 2002.

[40] J. Vreeken, "Spiking neural networks, an introduction," *technical report*, no. UU-CS-2003-008, Department of information and computing sciences, Utrecht University, 2003.

[41] G. Bi and M. Poo, "Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type," *J. Neurosci.*, vol. 18, pp. 10464–10472, 1998.

[42] S. Song, K. D. Miller, and L. F. Abbott, "Competitive hebbian learning through spike-timing dependent synaptic plasticity," *Nature Neuroscience*, vol. 3, no. 9, 919–926, 2000.

[43] G. Q. Bi and H. X. Wang, "Temporal asymmetry in spike timing-dependent plasticity," *Physiol. Behavior*, vol. 77, pp. 551–555, 2002.

[44] A. P. Shon, R. P. N. Rao, and T. J. Sejnowski, "Motion detection and prediction through spike-timing dependent plasticity," *Network: Comput. Neural Syst.*, vol. 15, no. 3, pp. 179–198, 2004.

[45] D. D. Coon and A. G. U. Perera, "Integrate-and-fire coding and Hodgkin-Huxley circuits employing silicon diodes," *Neural Networks*, vol. 2, no. 2, pp. 143–151, 1989.

[46] C. Mead, "Neuromorphic electronic systems," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990.

[47] M. Mahowald and C. Mead, "The silicon retina," *Scientific American*, vol. 264, no. 5, pp. 76–82, 1991.

[48] M. Mahowald and R. Douglas, "A silicon neuron," *Nature*, vol. 354, pp. 515–518, 1991.

[49] C. Song and K. P. Roenker, "Novel heterostructure device for electronic pulse-mode neural circuits," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 663–665, 1994.

[50] R. Douglas, M. Mahowald, and C. Mead, "Neuromorphic analogue VLSI," *Annual Review of Neuroscience*, vol. 18, pp. 255–281, 1995.

[51] C. Diorio and R. P. N. Rao, "Neural circuits in silicon," *Nature*, vol. 405, no. 6789, pp. 891–892, 2000.

[52] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, et. al., "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947–951, 2000.

[53] G. Indiveri, "A neuromorphic VLSI device for implementing 2-D selective attention systems," *IEEE Transactions on Neural Networks*, vol. 12, no. 6, pp. 1455–1463, 2001.

[54] C. Diorio, D. Hsu, and M. Figueroa, "Adaptive CMOS: from biological inspiration to systems-on-a-chip," *Proceedings of the IEEE*, vol. 90, no. 3, pp. 345–357, 2002.

[55] E. Chicca, D. Badoni, V. Dante, et al., "A VLSI recurrent network of integrate-and-fire neurons connected by plastic synapses with long-term memory," *IEEE Transactions on Neural Networks*, vol. 14, no. 5, pp. 1297–1307, 2003.

[56] K. A. Zaghloul and K. Boahen, "Optic nerve signals in a neuromorphic chip I: Outer and inner retina models," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 4, pp. 657–666, Apr. 2004.

[57] A. Bofill-i-Petit and A. F. Murray, "Synchrony detection and amplification by silicon neurons with STDP synapses," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1296–1304, 2004.

[58] S. Liu and R. Douglas, "Temporal coding in a silicon network of integrate-and-fire neurons," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1305–1314, 2004.

[59] R. J. Vogelstein, U. Mallik, J. T. Vogelstein, and G. Cauwenberghs, "Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses," *IEEE Trans. Neural Networks*, vol. 18, no. 1, pp. 253–265, 2007.

[60] K. A. Boahen, "Point-to-point connectivity between neuromorphic chips using address-events," *IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process.*, vol. 47, no. 5, pp. 416–434, May 2000.

[61] C. Rasche, "Neuromorphic excitable maps for visual processing," *IEEE Trans. Neural Networks.*, vol. 18, no. 2, pp. 520–529, 2007.

[62] P. A. Merolla, J. V. Arthur, B. E. Shi, and K. A. Boahen, "Expandable networks for neuromorphic chips," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 54, no. 2, pp. 301–311, 2007.

[63] B. Hille, *Ionic Channels of Excitable Membranes*. Sunderland, MA: Sinauer, 1992.

[64] E. Farquhar and P. Hasler, "A bio-physically inspired silicon neuron," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 52, no. 3, pp. 477–488, 2005.

[65] C. Koch, "Computation and the single neuron," *Nature*, vol. 358, no. 6613, pp. 207–210, 1997.

[66] T. Lehmann, "Hardware learning in analogue VLSI neural networks," PhD thesis, Technical University of Denmark, Denmark, 1994.

[67] S. Jones, R. Meddis, S. C. Lim, and A. R. Temple, "Toward a digital neuromorphic pitch extraction system," *IEEE Transaction on Neural Networks*, vol. 11, no. 4, pp. 978–987, 2000.

[68] T. Schoenauer, S. Atasoy, N. Mehrtash, and H. Klar, "NeuroPipe-chip: a digital neuro-processor for spiking neural networks," *IEEE Transaction on Neural Networks*, vol. 13, no. 1, pp. 205–213, 2002.

[69] J. N. H. Heemskerk, "Neurocomputers for brain-Style processing: design, implementation and application," PhD thesis, Leiden University, The Netherlands, 1995.

[70] M. R. DeYong, R. L. Findley, and C. Fields, "The design, fabrication, and test of a new VLSI hybrid analog-digital neural processing element," *IEEE Transaction on Neural Networks*, vol. 3, no. 3, pp. 363–374, 1992.

[71] Y. Horio, K. Aihara, and O. Yamamoto, "Neuron-synapse IC chip-set for large-scale chaotic neural networks," *IEEE Transaction on Neural Networks*, vol. 14, no. 5, pp. 1393–1404, 2003.

[72] T. Delbruck, "Silicon retina with correlation-based velocity-tuned pixels," *IEEE Trans. Neural Networks*, vol. 4, no. 3, pp. 529–541, 1993.

[73] C. Koch and B. Mathur, "Neuromorphic vision chips," *IEEE Spectrum*, vol. 33, pp. 38–64, 1996.

[74] K. Boahen, "Retinomorphic vision system," in *Proceedings of International Conference on Microelectronics for Neural Networks (MicroNeuro)*, Lausanne, 1996, pp. 2–14.

[75] Z. K. Kalayjian and A. G. Andreou, "A silicon retina for polarization contrast vision," in *Proc. Int. Joint Conf. on Neural Networks*, Washington, 1999, vol. 4, pp. 2329–2332.

[76] K. A. Zaghloul and K. Boahen, "Optic nerve signals in a neuromorphic chip II: Testing and results," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 4, pp. 667–675, Apr. 2004.

[77] R. F. Lyon and C. A. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 7, pp.1119–1134, Jul. 1988.

[78] E. Fragniere, A. van Schaik, and E. Vittoz, "Design of an analogue VLSI model of an active Cochlea," *Analog Integrated Circuits and Signal Processing*, vol. 13, no. 1-2, pp. 19–35, 1997.

[79] T. Horiuchi and K. M. Hynna, "A VLSI-based model of azimuthal echolocation in the big brown bat," *Auton. Robots*, vol. 11, no. 3, pp. 241–247, 2001.

[80] R. Sarpeshkar, J. Kramer, G. Indiveri, and C. Koch, "Analog VLSI architectures for motion processing: From fundamental limits to system applications," *Proc. IEEE*, vol. 84, pp. 969–987, 1996.

[81] T. K. Horiuchi and C. Koch, "Analog VLSI-based modeling of the primate oculomotor system," *Neural Comput.*, vol. 11, no. 1, pp. 243–265, 1999.

[82] R. R. Harrison and C. Koch, "A robust analog VLSI motion sensor based on the visual system of the fly," *Auton. Robots*, vol. 7, no. 3, pp. 211–224, 1999.

[83] R. R. Harrison, "A biologically inspired analog IC for visual collision detection," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 52, no. 11, pp. 2308–2318, 2005.

[84] J. C. Principe, V. G. Tavares, J. G. Harris, and W. J. Freeman, "Design and implementation of a biological realistic olfactory cortex in analog VLSI," *Proc. IEEE*, vol. 89, no. 7, pp. 1030–1051, 2001.

[85] T. J. Koickal, A. Hamilton, S. L. Tan, et. al., "Analog VLSI circuit implementation of an adaptive neuromorphic olfaction chip," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 54, no. 1, pp. 60–73, 2007.

[86] C. Rasche, *The Making of a Neuromorphic Visual System*. New York: Springer-Verlag, 2005.

[87] G. Indiveri, R. Murer, and J. Kramer, "Active vision using an analog VLSI model of selective attention," *IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process.*, vol. 48, no. 5, pp. 492–500, 2001.

[88] T. Y. W. Choi, B. E. Shi, and K. A. Boahen, "An on-off orientation selective address event representation image transceiver chip," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 51, no. 2, pp. 342–353, 2004.

[89] T. Y. W. Choi, P. A. Merolla, J. V. Arthur, K. A. Boahen, and B. E. Shi, "Neuromorphic implementation of orientation hypercolumns," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 52, no. 6, pp. 1049–1060, 2005.

[90] R. J. Vogelstein, U. Mallik, E. Culurciello, G. Cauwenberghs, and R. Etienne-Cummings, "A multichip neuromorphic system for spike-based visual information processing," *Neural Comput.*, vol. 19, no. 9, pp. 2281–2300, 2007.

[91] M. Sivilotti, "Wiring considerations in analog VLSI systems, with application to field-programmable networks," Ph.D. dissertation, Dept. Comp. Sci., California Inst. Technol., Pasadena, 1991.

[92] J. Lazzaro, J.Wawrzynek, M. Mahowald, M. Sivilotti, and D. Gillespie, "Silicon auditory processors as computer peripherals," *IEEE Trans. Neural Networks*, vol. 4, no. 3, pp. 523–528, 1993.

[93] M. Mahowald, *An Analog VLSI System for Stereoscopic Vision*. Norwell, MA: Kluwer, 1994.

[94] K. A. Boahen, "Communicating neuronal ensembles between neuromorphic chips," in *Neuromorphic Systems Engineering*, T. S. Lande, Ed. Norwell, MA: Kluwer, 1998, pp. 229–259.

[95] K. A. Boahen, "A burst-mode word-serial address-event link-I: Transmitter design," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 51, no. 7, pp. 1269–1280, 2004.

[96] —, "A burst-mode word-serial address-event link-II: Receiver design," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 51, no. 7, pp. 1281–1291, 2004.

[97] —, "A burst-mode word-serial address-event link-III: Analysis and test results," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 51, no. 7, pp. 1292–1300, 2004.

[98] D. Dupeyron, S. Le Masson, Y. Deval, G. Le Masson, and J. P. Dom, "A BiCMOS implementation of the Hodgkin-Huxley formalism," in *Proc. 5th Int. Conf. Microelectronics for Neural Networks and Fuzzy Systems*, pp. 311–316, 1996.

[99] J. Shin and C. Koch, "Dynamic range and sensitivity adaptation in a silicon spiking neuron," *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 1232–1238, 1999.

[100] C. Rasche and R. Douglas, "An improved silicon neuron," *Analog Integrated Circuits and Signal Processing*, vol. 23, pp. 227–236, 2000.

[101] M. F. Simoni, G. S. Cymbulyuk, M. E. Sorensen, R. L. Calabrese, and S. P. DeWeerth, "A multiconductance silicon neuron with biologically matched dynamics," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 342–354, 2004.

[102] J. V. Arthur and K. Boahen, "Synchrony in silicon: the gamma rhythm," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1815–1825, 2007.

[103] S. R. Schultz and M. A. Jabri, "Analogue VLSI integrate-and-fire neuron with frequency adaptation," *Electron. Lett.*, vol. 31, pp. 1357–1358, 1995.

[104] A. van Schaik, "Building blocks for electronic spiking neural networks," *Neural Networks*, vol. 14, pp. 617–628, 2001.

[105] G. Indiveri, E. Chicca, and R. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 211–221, 2006.

[106] S. Fusi, M. Annunziato, D. Badoni, A. Salamon, and D. J. Amit, "Spike-driven synaptic plasticity: theory, simulation, VLSI implementation," *Neural Computation*, vol. 12, 2227–2258, 2000.

[107] M. E. Zaghloul, J. L. Meador, and R. W. Newcomb, *Silicon Implementation of Pulse Coded Neural Networks*. Springer, 1994, pp. 153–164.

[108] J. V. Arthur and K. Boahen, "Recurrently connected silicon neurons with active dendrites for one-shot learning," in *Proc. Int. Joint Conf. on Neural Networks*, 2004, vol. 3, pp. 1699–1704.

[109] P. Merolla and K. Boahen, "A recurrent model of orientation maps with simple and complex cells," in *Advances in Neural Information Processing Systems*, vol. 16, pp. 995–1002, MIT Press, 2003.

[110] K. Hynna and K. Boahen, "Space-rate coding in an adaptive silicon neuron," *Neural Networks*, vol. 14, pp. 645–656, 2001.

[111] G. Indiveri, "Modeling selective attention using a neuromorphic analog VLSI device," *Neural Computation*, vol. 12, no. 12, pp. 2857–2880, 2000.

[112] S. C. Liu, J. Kramer, G. Indiveri, et. al., "Orientation-selective aVLSI spiking neurons," *Neural Networks*, vol. 14, no. 6/7, pp. 629–643, 2001.

[113] C. Rasche and R. Hahnloser, "Silicon synaptic depression," *Biological Cybernetics*, vol. 84, no. 1, pp. 57–62, 2001.

[114] M. Boegerhausen, P. Suter, and S. C. Liu, "Modeling short-term synaptic depression in silicon," *Neural Computation*, vol. 15, no. 2, pp. 331–348, 2003.

[115] A. Bofill, A. Murray, and D. Thompson, "Circuits for VLSI implementation of temporally-asymmetric hebbian learning," in *Advances in Neural Information Processing Systems*, vol. 14, Cambridge, MA: MIT Press, 2002.

[116] G. Indiveri, "Neuromorphic bistable VLSI synapses with spike timing dependent plasticity," in *Advances in Neural Information Processing Systems*, vol. 15, Cambridge, MA: MIT Press, 2002.

[117] J. V. Arthur and K. Boahen, "Learning in silicon: timing is everything," in *Advances in Neural Information Processing Systems*, vol. 18, MIT Press, 2006.

[118] S. Mitra, S. Fusi, and G. Indiveri, "A VLSI spike-driven dynamic synapse which learns only when necessary," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, 2006, pp. 2777–2780.

[119] E. Chicca, G. Indiveri, and R. Douglas, "An adaptive silicon synapse," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, 2003, pp. 81–84.

[120] P. Hasler, C. Diorio, B. A. Minch, and C. Mead, "Single transistor learning synapses with long term storage," *IEEE Int. Symp. on Circuits and Systems*, vol. 3, pp. 1660-1663, 1995.

[121] C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A single-transistor silicon synapse," *IEEE Trans. Electron Devices*, vol. 43, no.11, pp. 1972–1980, 1996.

[122] C. Diorio, D. Hsu, and M. Figueroa, "Adaptive CMOS: from biological inspiration to systems-on-a-chip," *Proceedings of the IEEE*, vol. 90, no. 3, pp. 345–357, 2002.

[123] C. Diorio, "A p-channel MOS synapse transistor with self-convergent memory writes," *IEEE Trans. Electron Devices*, vol. 47, no.2, pp. 464–472, 2000.

[124] C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A floating-gate MOS learning array with locally computed weight updates," *IEEE Trans. Electron Devices*, vol. 44, no.12, pp. 2281–2289, 1997.

[125] P. Hasler and J. Dugger, "Correlation learning rule in floating-gate pFET synapses," *IEEE Trans. Circuits and Systems II*, vol. 48, no.1, pp. 65–73, 2001.

[126] D. Hsu, M. Figueroa, and C. Diorio, "Competitive learning with floating-gate circuits," *IEEE Trans. Neural Networks*, vol. 13, no.3, pp. 732–744, 2002.

CHAPTER 2 FUNDAMENTALS OF SEMICONDUCTOR DEVICES

Section 2.1 Introduction

The semiconductor industry has been one of the most successful on the planet in the past few decades. Semiconductor devices are electronic components that exploit the electronic properties of semiconductor materials. Integrated circuits (ICs), which consist of a number (from a few to millions) of these devices manufactured and interconnected on a single semiconductor substrate, are found inside all modern electronic appliances and products. Since the invention of the transistor in 1950s, manufacturers have been successful in developing ever more complex ICs by making the individual transistors smaller and finding ways to combine more of them together on a single chip. There has been a regular increase in the computational and processing capability of ICs by doubling the number of transistors every two years. The fast progress of semiconductor technology provides powerful resources for producing brain-like intelligence in VLSI, and the principles of semiconductor devices form the basis of the implementation techniques presented in this thesis.

This chapter is intended to provide fundamentals of semiconductor devices which are relevant to the work presented in the thesis. In Section 2.2, the principles and electric characteristics of the Metal-Oxide-Semiconductor (MOS) capacitor are described, together with an experimental study to support the discussion. A brief explanation of the operation of the MOS transistor under dc conditions is provided in Section 2.3. The Silvaco software package is employed to simulate the MOS transistor. Four of the interactive tools will be used: (1) DECKBUILD, the command center of Silvaco's interactive simulation environment. It provides convenient way to specify problems, run simulations, switch between simulators, extract parameters from simulation results, and invoke other Virtual Wafer Fab (VWF) interactive tools. (2) DEVEDIT, a device structure editor, can be used to generate a new mesh on an existing structure, modify a device or create a device from scratch. These devices can then be used by Silvaco 2-D and 3-D simulators. (3) TONYPLOT, a graphical post processing tool for

use with all Silvaco simulators. It can operate stand-alone, or along with other VWF interactive tools such as Deckbuild. (4) ATLAS, a physically based two and three dimensional device simulator. The simulation study will predict the electrical behavior of specified semiconductor structures and provides insight into the internal physical mechanisms associated with device operation. Conclusions of this chapter are given in Section 2.4.

Section 2.2 MOS Capacitor

One of the important building blocks of semiconductor devices is the MOS capacitor. As shown in Fig. 2.1, it consists of a metal/polysilicon contact, referred to as the gate, on top of a thin oxide layer grown on a semiconductor substrate.

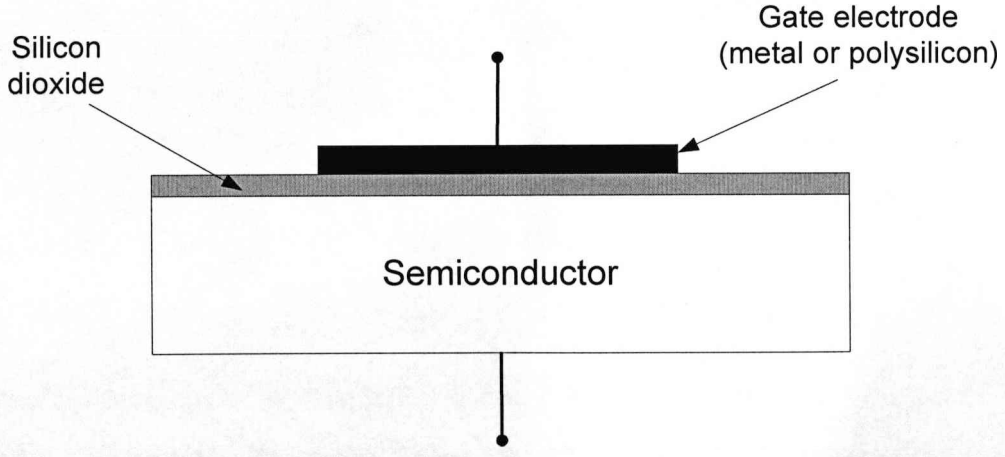


Fig. 2.1 Schematic structure of a MOS capacitor.

The energy bands for each material of a *p*-type MOS capacitor are shown in Fig. 2.2. In an ideal MOS capacitor, the metal work function Φ_m, is equal to the semiconductor work function Φ_{si}. Therefore, the Fermi level of the semiconductor E_F is aligned with the Fermi level of the gate, and there is no band bending in any region of the MOS capacitor. In addition, the gate dielectric is assumed to be free of any charges and the semiconductor is uniformly doped.

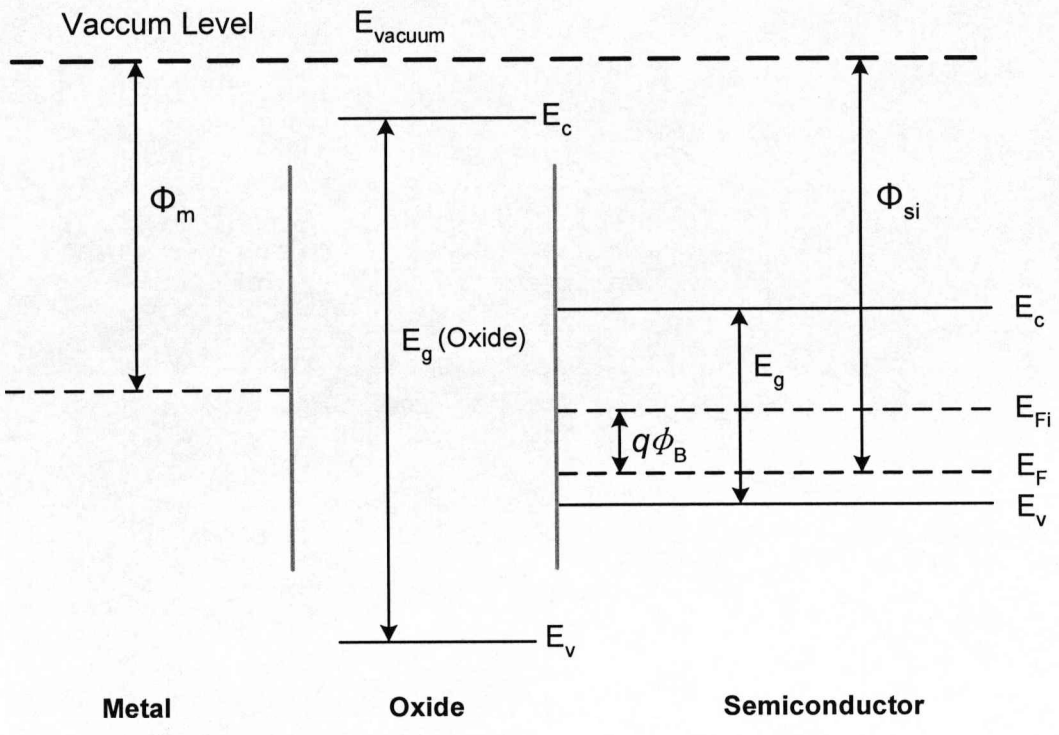


Fig. 2.2 Energy band diagram in each material when separate.

Section 2.2.1 Basic Operations

As shown in Fig. 2.3, there are three different modes of operation in a MOS capacitor: accumulation, depletion and inversion, each of which is under a certain bias voltage, one below the flatband voltage V_{FB}, one between the flatband voltage and the threshold voltage V_T, and one larger than the threshold voltage [1].

Accumulation occurs for the voltages less than the flatband voltage for a *p*-type substrate. In a MOS capacitor charge neutrality is always preserved. Therefore net positive charge Q_{si} must be provided in the silicon substrate to counterbalance the negative charge on the gate. This is achieved by an accumulation of majority carrier holes under the gate. Therefore, the majority carrier concentration is greater near the oxide-semiconductor interface compared to the bulk. Under an applied negative gate bias, the Fermi level of the gate is raised with respect to the Fermi level of the

substrate by an amount equal to qV_g, as shown in Fig. 2.4(a). While the Fermi level in the substrate remains invariant even under an applied bias since no current can flow through the device due to the presence of an insulator. The energy bands in the semiconductor bend upward bringing the valence band closer to the Fermi level indicative of a higher hole concentration under the dielectric.

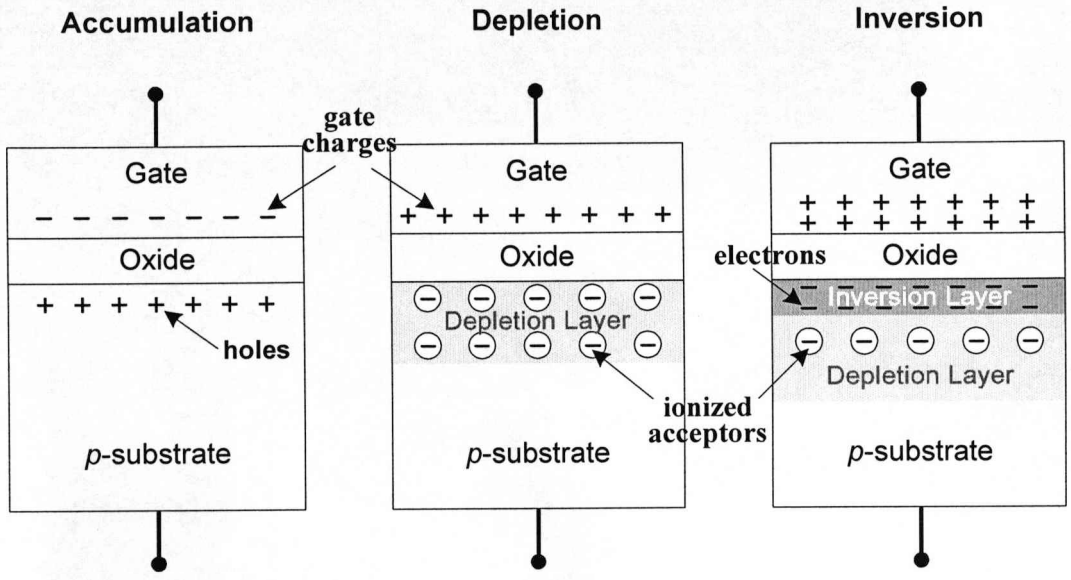


Fig. 2.3 Schematic view of the charge distributions in a MOS capacitor under different conditions: accumulation, depletion and inversion.

Depletion occurs for positive voltages for a *p*-type substrate. The mobile holes are repelled from the oxide-semiconductor interface and the ionized acceptor ions are exposed in the space charge region. The charge in the depletion region is exactly equal to the charge on the gate in order to preserve charge neutrality. With a positive gate bias, the Fermi level of the gate is lowered with respect to the Fermi level of the substrate [2]. As shown in Fig. 2.4(b), the bands bend downward resulting in a positive surface potential ϕ_s. Under the gate, the valence band moves away from the Fermi level indicative of hole depletion. When the band bending at the surface is such that the intrinsic level coincides with the Fermi level, the surface resembles an intrinsic material. The surface potential required to have this condition is given by:

$$\phi_s = \phi_B = \frac{1}{q}(E_{Fi} - E_F) = V_t \ln \frac{N_a}{n_i} \tag{2.1}$$

where ϕ_B is the bulk potential; E_{Fi} is the intrinsic Fermi level; E_F is the Fermi level of the semiconductor; V_t is the thermal voltage; N_a is the acceptor density in the substrate; n_i is the intrinsic concentration.

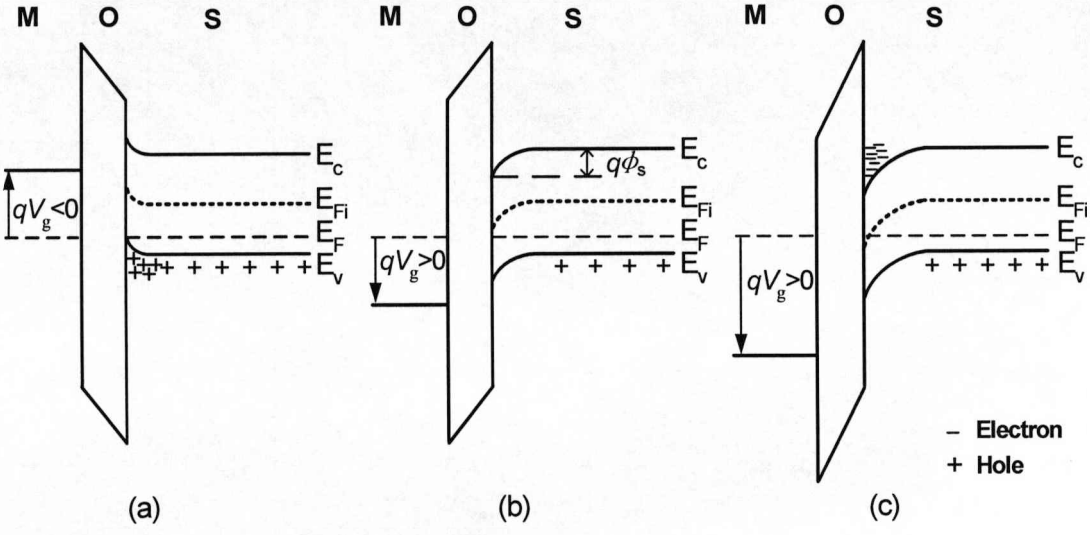


Fig. 2.4 Energy band diagrams in a *p*-type MOS capacitor under accumulation, depletion, and inversion.

In the depletion layer, the total charge per unit area Q_d can be described as:

$$Q_d = -qN_aW_d \tag{2.2}$$

where W_d is the depletion layer depth.

Assuming negligible free charge at the interface, integration of the charge density yields the electric field in the semiconductor at the interface ξ_{si}, and the field in the oxide ξ_{ox}:

$$\xi_{si} = \frac{qN_aW_d}{\varepsilon_{si}\varepsilon_0} \tag{2.3}$$

$$\xi_{ox} = \frac{qN_aW_d}{\varepsilon_{ox}\varepsilon_0} \tag{2.4}$$

The electric field in the semiconductor changes linearly and is zero at the edge of the depletion region. The surface potential is obtained by integrating the electric field with respect to W_d:

$$\phi_s = \int \xi_{si} dW_d = \frac{qN_a W_d^2}{2\varepsilon_{si}\varepsilon_0} \tag{2.5}$$

When the surface potential equals twice the bulk potential, the electron density at the surface equals N_a. This corresponds to the situation where gate voltage is at threshold of inversion operation. So the depletion depth is restricted to the range:

$$W_d = \sqrt{\frac{2\varepsilon_{si}\varepsilon_0\phi_s}{qN_a}}, \quad 0 \le \phi_s \le 2\phi_B \tag{2.6}$$

The relation between the potential across the depletion region and the gate voltage is given by the following equation which can be used to derive the capacitance corresponding to the gate voltage:

$$V_g = V_{FB} + \phi_s - \frac{Q_d}{C_{ox}} \tag{2.7}$$

Substituting (2.2) and (2.6) into (2.7) gives:

$$V_g = V_{FB} + \phi_s + \frac{\sqrt{2\varepsilon_{si}\varepsilon_0\phi_s qN_a}}{C_{ox}}, \quad 0 \le \phi_s \le 2\phi_B \tag{2.8}$$

Inversion occurs when voltages exceed the threshold voltage V_T. Under a larger positive gate bias, the positive charge on the gate increases further and the oxide field begins to collect thermally generated electrons under the gate. The band bending also increases so that the intrinsic energy level at the surface goes lower than the Fermi level, as shown in Fig 2.4(c). With the presence of minority carrier electrons, the intrinsic surface therefore begins to change into an *n*-type inversion layer. The negative charge in the semiconductor is comprised of ionized acceptor atoms in the depletion region and free electrons in the inversion layer. At this point, the electron concentration at the surface is still less than the hole concentration in the neutral bulk. Thus, this condition is referred to as weak inversion. As the gate bias is increased further, the band bending increases. The depletion region becomes wider and the electron concentration in the inversion layer increases. When the electron

concentration is equal to the hole concentration in the bulk, a strong inversion layer is formed. The surface potential required for the onset of strong inversion is:

$$\phi_s = 2\phi_B = 2V_t \ln\frac{N_a}{n_i} \tag{2.9}$$

So the depletion layer width reaches its maximum value W_{dm}:

$$W_{dm} = \sqrt{\frac{4\varepsilon_{si}\varepsilon_0\phi_B}{qN_a}} \tag{2.10}$$

and the expression for the gate voltage V_g is:

$$V_g = V_{FB} + \phi_s - \frac{Q_d + Q_{inv}}{C_{ox}} = V_{FB} + \phi_s + \frac{\sqrt{4\varepsilon_{si}\varepsilon_0\phi_B qN_a}}{C_{ox}} - \frac{Q_{inv}}{C_{ox}} \tag{2.11}$$

where Q_{inv} is the charge in inversion layer.

The relationship among the surface potential, electric field, and charge can be derived by solving Poisson's equation in the surface region of silicon [3]. By using Gauss's law, the total charge per unit area in the silicon is given as:

$$Q_{si} = \pm\sqrt{2\varepsilon_{si}\varepsilon_0 V_t qN_a}\left[\left(e^{-\phi_s/V_t} + \frac{\phi_s}{V_t} - 1\right) + \frac{n_i^2}{N_a^2}\left(e^{\phi_s/V_t} - \frac{\phi_s}{V_t} - 1\right)\right]^{1/2} \tag{2.12}$$

At the accumulation condition, $\phi_s<0$ and the first term in the square bracket dominates when $-\phi_s/V_t>1$. Therefore the accumulation charge density is proportional to $\exp(-\phi_s/2V_t)$. In depletion, $\phi_s > 0$ and $\phi_s/V_t > 1$. Since the term $(n_i^2/N_a^2)\exp(\phi_s/V_t)$ is not large enough, ϕ_s/V_t in the first term dominates and the negative depletion charge is proportional to $\phi_s^{1/2}$. When the inversion occurs, the $(n_i^2/N_a^2)\exp(\phi_s/V_t)$ term becomes larger than the ϕ_s/V_t term. The negative inversion charge density is then proportional to $\exp(\phi_s/2V_t)$.

Section 2.2.2 Capacitance-Voltage Characteristics

The capacitance-voltage (C-V) response is the essential technique to characterize the MOS system. The MOS structure is considered as a series connection of two capacitors: the oxide capacitance C_{ox} and the capacitance of the substrate C_{si}.

In the accumulation mode of the MOS capacitor, a layer of charges is formed by majority carriers. Thus the MOS capacitor can be treated as a series connection of oxide capacitance and accumulation capacitance:

$$\frac{1}{C}=\frac{1}{C_{ox}}+\frac{1}{C_{si}} \tag{2.13}$$

$$C_{si}=\frac{dQ_{si}}{d\phi_s} \tag{2.14}$$

Since the accumulation charge Q_{si} is exponentially dependent on ϕ_s, C_{si} becomes very large for negative values of V_g in the case of p-type substrate. Therefore the effect of Q_{si} can be ignored, and only the value of C_{ox} will be observed in the accumulation region of a C-V plot for both low frequency and high frequency:

$$C_{LF}=C_{HF}=C_{ox}, \text{ for } V_g \le V_{FB} \tag{2.15}$$

In depletion mode, the majority carriers are pushed away from the oxide-semiconductor interface and a depletion region is formed. The MOS capacitance is obtained from the series connection of the oxide capacitance and the capacitance of the depletion layer:

$$\frac{1}{C_{LF}}=\frac{1}{C_{HF}}=\frac{1}{C_{ox}}+\frac{1}{C_d} \tag{2.16}$$

$$C_d=\frac{\varepsilon_{si}\varepsilon_0}{W_d} \tag{2.17}$$

where W_d is the depletion layer width given by (2.6). With the increasing of the gate voltage, the depletion width increases. Therefore the depletion capacitance C_d decreases and so does the total MOS capacitance.

In inversion mode, the total capacitance does not depend on the gate voltage. In the low frequency case, minority carrier generation can follow the gate signal and the inversion charge Q_{inv} increase exponentially with gate voltage V_g. Therefore the total capacitance in the substrate is much larger than the oxide capacitance C_{ox}. As a result, the low frequency capacitance equals the oxide capacitance:

$$C_{LF}=C_{ox}, \text{ for } V_g \ge V_T \tag{2.18}$$

In the case of high frequency (such as 1MHz), the minority carrier generation cannot follow the ac signal, but can follow the dc bias. Thus only the depletion capacitance affects the measured capacitance and the MOS capacitance is obtained from:

$$\frac{1}{C_{HF}} = \frac{1}{C_{ox}} + \frac{1}{C_{dm}}, \text{ for } V_g \geq V_T \tag{2.19}$$

$$C_{dm} = \frac{\varepsilon_{si}\varepsilon_0}{W_{dm}} \tag{2.20}$$

where W_{dm} is the depletion layer width at threshold, given by (2.10).

Section 2.2.3 Deep Depletion Capacitance

When the gate voltage is ramped very fast from flatband to threshold and beyond, the MOS capacitor operates in the thermal non-equilibrium state, namely deep depletion. The majority carriers are repelled over the depth of the depleted space charge region in a short time, the dielectric relaxation time $\tau_{relax} = \varepsilon_{si}\varepsilon_0/\sigma$ where σ is conductivity and $\varepsilon_{si}\varepsilon_0$ is permittivity. The value for τ_{relax} is about 10^{-10}s for the doping levels used in this work. The depletion layer width then goes beyond its maximum equilibrium value, and the inversion layer is not formed at this time. Consequently the device is restored to equilibrium through thermal generation of electron-hole pairs, provided that there are no other sources of minority carriers [4].

The capacitance depends on the gate voltage and the inversion charge Q_{inv} as:

$$C(t) = \frac{C_{ox}}{\sqrt{1 + \dfrac{2(V_g - V_{FB} + Q_{inv}(t)/C_{ox})}{V_0}}} \tag{2.21}$$

where $V_0 = q\varepsilon_{si}\varepsilon_0 N_a / C_{ox}^2$. Solving (2.21) and differentiating with respect to t yields:

$$\frac{dV_g}{dt} = -\frac{1}{C_{ox}}\frac{dQ_{inv}}{dt} - \frac{q\varepsilon_{si}\varepsilon_0 N_a}{C^3}\frac{dC}{dt} \tag{2.22}$$

Since V_g is constant for the pulsed MOS capacitor, $dV_g/dt = 0$ and the thermal generation rate can be derived from (2.22) as:

$$\frac{dQ_{inv}}{dt} = \frac{q\varepsilon_{si}\varepsilon_0 N_a}{C^3}\frac{dC}{dt} \tag{2.23}$$

In MOS capacitors, the occurrence of deep depletion is linked to the carrier lifetime. Therefore the capacitance-time measuring technique is generally used to determine either recombination or generation lifetime [4]. The basic method proposed by Zerbst in 1966 involves a 'Zerbst plot' [5], where the vertical axis is proportional to the generation current and the horizontal axis is proportional to the generation width, easily extracted from the measured capacitance. It takes a few seconds for MOS capacitors with a long lifetime to relax to thermal equilibrium, whereas relaxation time of the order of milliseconds may be required for the MOS capacitors with a short lifetime. Further theoretical analysis and simulation study are presented in the context of the solid-state synapse in Chapter 4.

Section 2.2.4 C-V Measurement and Parameter Extraction

The C-V analysis is illustrated by a short experimental study. The measured MOS capacitor was fabricated on *n*-type wafer. The diameter of the gate area is 2mm. The experiments were carried out on the C-V rig in the clean room. The high frequency characteristic is measured by superimposing a small ac signal on the dc gate voltage. The gate bias from -5V to 5V with 0.05V step is set at the frequency of 1MHz. The experimental result is shown in Fig. 2.5. The maximum value of capacitance $C_{\max}$ which is in the accumulation region of a C-V plot is 1.696×10^3 pF, and the minimum value 273 pF is in the inversion region.

(A) Extraction of oxide thickness

The measurement enables the oxide thickness to be extracted from:

$$C_{ox} = \frac{C_{\max}}{A} = \frac{\varepsilon_{ox}\varepsilon_0}{t_{ox}} \tag{2.24}$$

where C_{ox} is the capacitance per unit area and A is the capacitor area. Substituting the experimental values and parameters, the oxide thickness is obtained: t_{ox}=64nm.

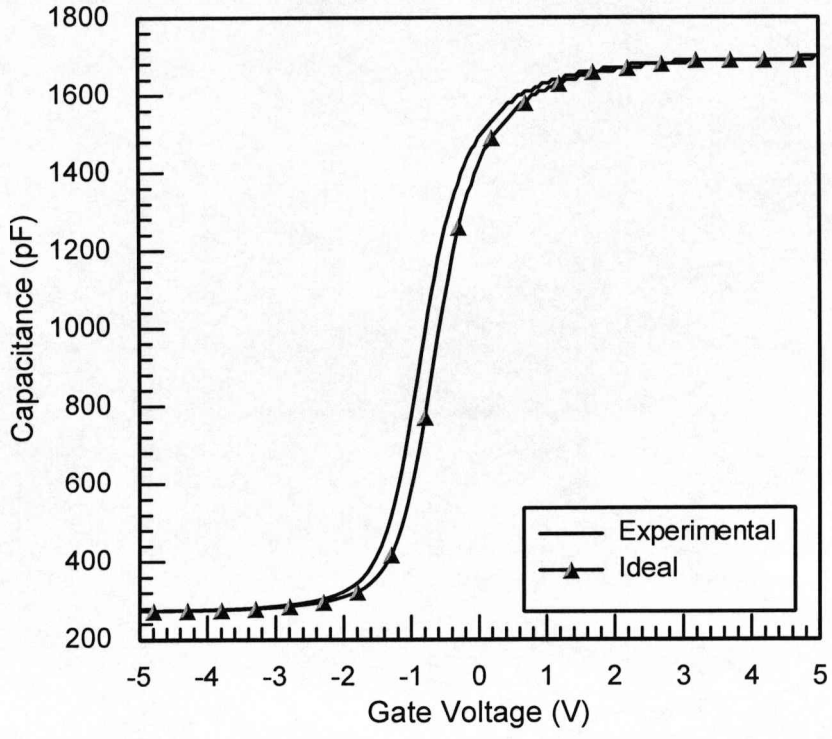


Fig. 2.5 C-V characteristic of the MOS capacitor with 2mm gate diameter.

(B) Extraction of doping density

In inversion, the minimum capacitance is given by the series connection of the oxide capacitance and the capacitance of the depletion layer. Since the oxide capacitance equals to the maximum measured capacitance, the minimum capacitance is derived:

$$\frac{1}{C_{\min}} = \frac{1}{C_{\max}} + \frac{1}{AC_d} \tag{2.25}$$

Substituting (2.20) and (2.10) into (2.25) gives:

$$\frac{1}{C_{\min}} = \frac{1}{C_{\max}} + \frac{1}{A}\sqrt{\frac{4\phi_B}{\varepsilon_{si}\varepsilon_0 q N_d}} \tag{2.26}$$

Hence, by solving the above equation the doping density N_d=2.49×10^{15}cm^{-3} was obtained.

(C) Extraction of flatband capacitance

The work function of an aluminum gate Φ_m is 4.1eV. The work function of an *n*-type semiconductor Φ_{si} is:

$$\Phi_{si} = \chi + \frac{E_g}{2} - qV_t \ln\frac{N_d}{n_i} = 4.3eV \tag{2.27}$$

Thus, the flatband voltage is:

$$V_{FB} = \frac{\Phi_m - \Phi_{si}}{q} = -0.2V \tag{2.28}$$

The ideal C-V curve with $V_{FB} = 0$ is plot in Fig. 2.5, and the measured capacitance at flatband voltage is: C_{FB}=1420 pF.

(D) Extraction of oxide charge

The oxide is a high quality insulator, however it contains both positive and negative electric charges in practice. These parasitic oxide charges cause simple linear shifts of the C-V characteristics along the voltage bias, but the overall shape is not distorted since oxide charges are not bias dependent. The positive charge causes the characteristics to be shifted to the left, while negative charge results in right-shift. The positive charges are typically positioned close to the oxide-semiconductor interface of the MOS capacitor, whereas negative charge is usually due to electrons trapped within the bulk of the oxide. The net oxide charge density is therefore expressed as:

$$n_{ox} = \frac{Q_{ox}}{q} = \frac{C_{ox}\Delta V}{q} = \frac{C_{\max}\Delta V}{qA} \tag{2.29}$$

where ΔV is the shift of the C-V plot, usually taken at the mid-gap point.

At mid-gap, the surface potential ϕ_s calculated by (2.1) is 0.31V, and the mid-gap capacitance C_{mg} obtained from (2.16) is 1.76×10^{-4} F/cm^2. Therefore, in the experiment the applied gate voltage $V_{g,experiment}$ at C_{mg} is -1.25V. The ideal gate voltage $V_{g,ideal}$ at C_{mg} can be calculated by (2.8). Hence, the oxide charge density is:

$$n_{ox} = \frac{\varepsilon_{ox}\varepsilon_0(V_{g,ideal} - V_{g,experiment})}{t_{ox}q} = 6.28\times10^{11}cm^{-2} \tag{2.30}$$

It is worth noting that the analysis given here is relevant to the shift in threshold voltage associated with charge stored on a floating gate; a concept that is relevant to the charge coupled synapses to be presented later.

(E) Surface state effect

Surface states are the electrically active states resulting from the disruption of the periodicity of the lattice at the semiconductor surface. These states within the band gap are typically donor-like in the lower part of the band gap and acceptor-like in the upper half. Donor-like states are neutral when occupied with electrons and positive when empty. Acceptor-like states are negative when occupied with electrons and neutral when empty of electrons. These states will result in an increasing right-shift of the C-V plot for gate voltages above mid-gap and left-shift below mid-gap, as the gate voltage is swept from the accumulation region to depletion and inversion region. The surface states affect the measured capacitance only at low frequency as there is time for their occupancy to follow the ac signal.

Section 2.3 MOS Transistor

The MOS transistor is at the core of the VLSI technology. The basic structure of an *n*-channel MOS transistor is shown in Fig. 2.6. It differs from the MOS capacitor in that two N+ regions (source and drain) are formed into the *p*-type substrate. Therefore in inversion the minority carriers are immediately supplied by the source and drain. The channel is the surface region under the gate oxide between source and drain, and is critical for current conduction in a MOS transistor. The mobile carriers can follow the gate-source voltage. When the gate-source voltage is zero, the *p*-type surface is either in accumulation or depletion and no current flow in the channel. When the gate-source voltage V_{gs} exceeds the threshold voltage V_T an inverted channel is formed, and an electron current will flow in the channel if there is a voltage difference between source and drain.

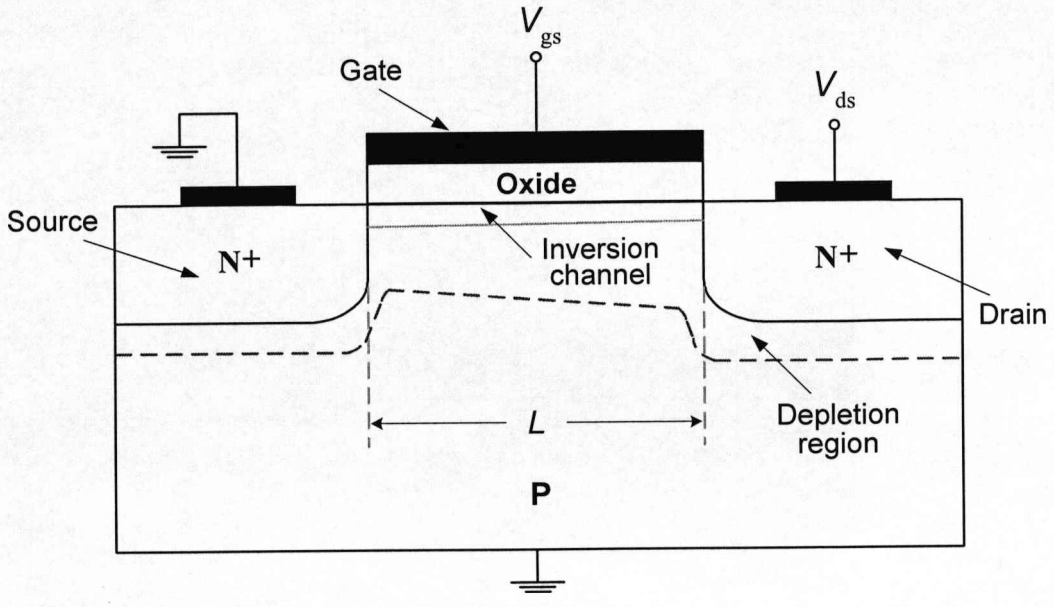


Fig. 2.6 A schematic cross section of MOS transistor, operated in the linear region (low V_{ds}).

Section 2.3.1 I-V Characteristics

In order to derive an analytical solution for the drain current, the inversion layer thickness is assumed to be zero. The surface potential for inversion condition is:

$$\phi_s = 2\phi_B + V(y) \tag{2.31}$$

where $V(y)$ is the electron quasi-Fermi potential along the channel with respect to the Fermi potential of the source. The depletion charge density is then:

$$Q_d = -qN_a W_{dm} = -\sqrt{2\varepsilon_{si}\varepsilon_0 qN_a(2\phi_B + V)} \tag{2.32}$$

The inversion charge density is:

$$Q_{inv} = Q_{si} - Q_d = -C_{ox}(V_{gs} - 2\phi_B - V - V_{FB}) + \sqrt{2\varepsilon_{si}\varepsilon_0 qN_a(2\phi_B + V)} \tag{2.33}$$

The drain current can be expressed as:

$$I_{ds} = \mu\frac{W}{L}\int_0^{V_{ds}}[-Q_{inv}(V)]dV \tag{2.34}$$

where μ is the mobility, W is the channel width, and L is the channel length.

Substituting (2.33) into (2.34) and carrying out the integration give:

$$I_{ds} = \mu C_{ox} \frac{W}{L}\left[(V_{gs} - 2\phi_B - \frac{V_{ds}}{2} - V_{FB})V_{ds} - \frac{2\sqrt{2\varepsilon_{si}\varepsilon_0 qN_a}}{3C_{ox}}[(2\phi_B + V_{ds})^{3/2} - (2\phi_B)^{3/2}]\right] \quad (2.35)$$

The above equation indicates the basic I-V characteristics of a MOS transistor. For a given V_{gs}, the current I_{ds} first increases linearly with the drain voltage V_{ds}, then levels off to a saturated value [3].

Section 2.3.2 Linear Operation

When $V_{ds} < V_{gs}$-V_T, the MOS transistor operates in the so-called linear region and (2.35) can be simplified as:

$$I_{ds} = \mu C_{ox} \frac{W}{L}\left[(V_{gs} - V_T)V_{ds} - \frac{m}{2}V_{ds}^2\right] \quad (2.36)$$

where m is the gate channel coupling:

$$m = 1 + \frac{C_{dm}}{C_{ox}} \quad (2.37)$$

and C_{dm} is the depletion capacitance at threshold, given by (2.20). V_T is the threshold voltage of the transistor given by:

$$V_T = 2\phi_B + \frac{\sqrt{4\varepsilon_{si}\varepsilon_0 qN_a\phi_B}}{C_{ox}} + V_{FB} \quad (2.38)$$

The drain current as a function of gate voltage at low drain voltage is obtained by Silvaco simulation and shown in Fig. 2.7. The threshold voltage V_T can be approximated from the extrapolated intercept of the linear portion of the curve.

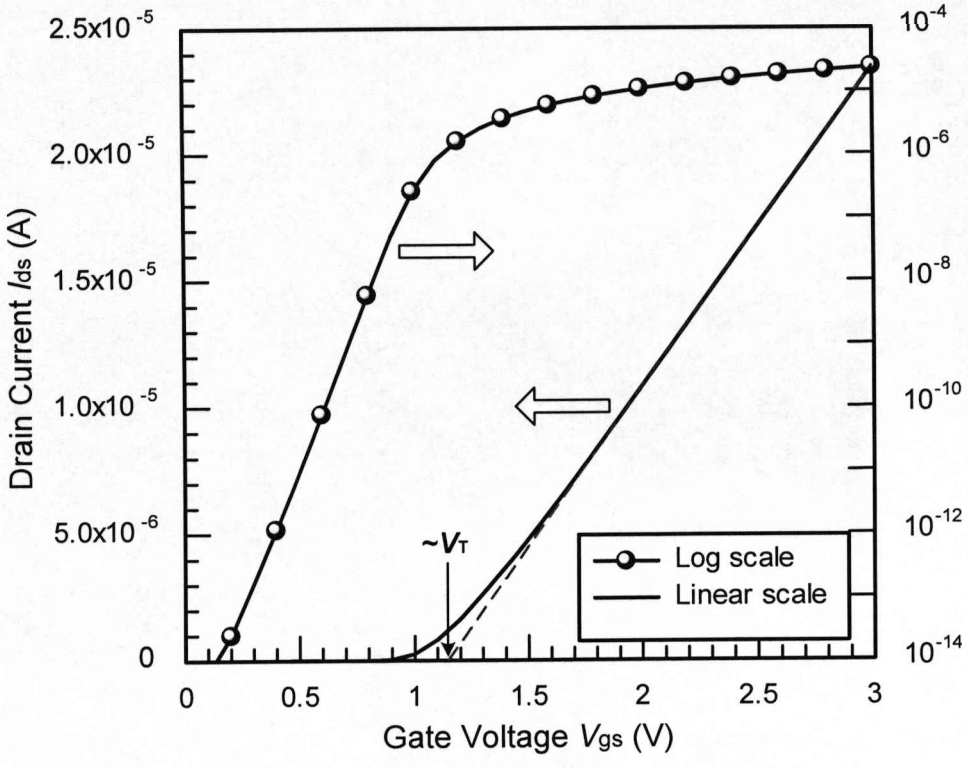


Fig. 2.7 I_{ds}-V_{gs} characteristic of a MOS transistor at low drain voltage (V_{ds}=0.1V), simulated in Silvaco. N_a= 2.12×10^{17} cm^{-3}; t_{ox}=16nm. V_T is determined by the linearly extrapolated intercept.

Section 2.3.3 Saturation Operation

As shown in Fig. 2.8, with the increase of V_{ds} the drain current I_{ds} follows a parabolic curve until a saturation value is reached:

$$I_{ds} = I_{sat} = \mu C_{ox} \frac{W}{L} \frac{(V_{gs} - V_T)^2}{2m} \tag{2.39}$$

Note that the depletion charge is neglected in this case. Beyond the saturation point, the increase of the current is due to the reduction of the effective channel length with increased V_{ds}. Therefore the saturation current is usually modeled by multiplying (2.39) by the factor $(1+\lambda V_{ds})$, where λ is the channel length modulation factor.

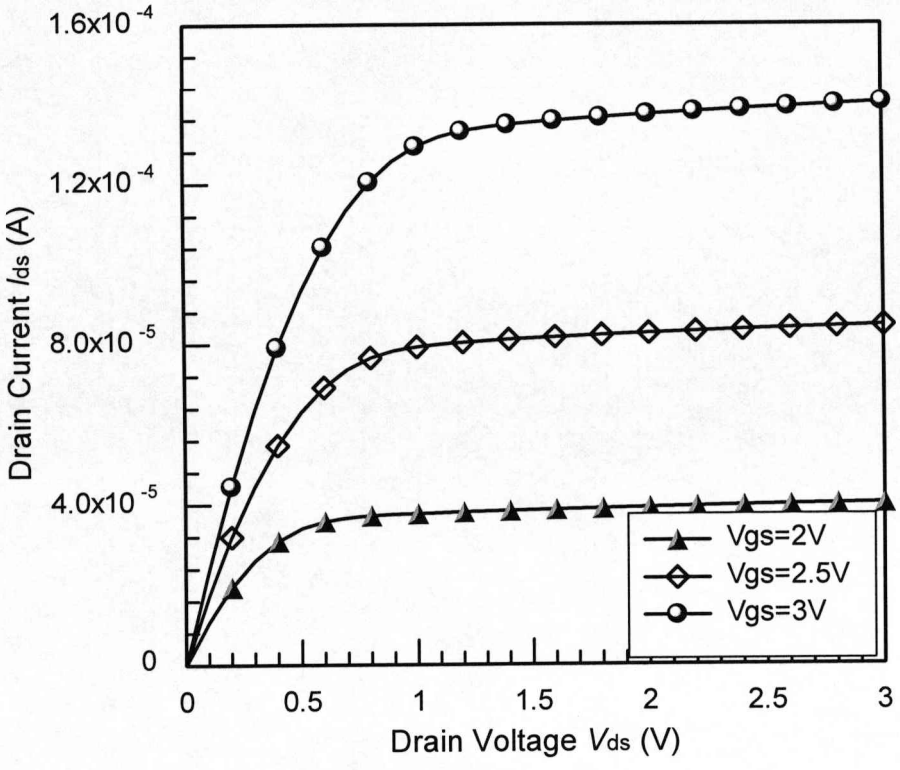


Fig. 2.8 I_{ds}-V_{ds} characteristic of a long channel MOS transistor for different gate-source voltages. Physical parameters in Silvaco: N_a= 2.12×10^{17} cm^{-3}; t_{ox}=16nm.

Section 2.3.4 Subthreshold Operation

When V_{gs} is below V_T, the MOS transistor is said to be in the subthreshold region. The inversion charge density follows an exponential dependence on ϕ_s, and the channel conductance is dominated by the diffusion current. From (2.12), the charge density in silicon is obtained by keeping only the significant terms in the square bracket:

$$Q_{si} = -\sqrt{2\varepsilon_{si}\varepsilon_0 V_t q N_a}\left[\frac{\phi_s}{V_t} + \frac{n_i^2}{N_a^2}e^{(\phi_s - V)/V_t}\right]^{1/2} \tag{2.40}$$

Expand (2.40) into a power series, the first order term gives the inversion charge density:

$$Q_{inv} = -V_t \frac{n_i^2}{N_a^2} e^{(\phi_s - V)/V_t}\sqrt{\frac{\varepsilon_{si}\varepsilon_0 q N_a}{2\phi_s}} \tag{2.41}$$

The drain current in the subthreshold region is obtained by substituting Q_{inv} into (2.34) and carrying out the integration:

$$I_{ds} = \mu \frac{W}{L} \frac{V_t^2 n_i^2}{N_a^2} e^{\phi_s / V_t} (1 - e^{-V_{ds}/V_t}) \sqrt{\frac{\varepsilon_{si} \varepsilon_0 q N_a}{2\phi_s}} \tag{2.42}$$

where ϕ_s can be expressed as a function of gate-source voltage using (2.8). If we assume that ϕ_s is slightly deviated from $2\phi_B$ [6], (2.8) is simplified as:

$$V_{gs} = V_t + m(\phi_s - 2\phi_B) \tag{2.43}$$

Substituting (2.43) into (2.42) yields the subthreshold current as a function of V_{gs}:

$$I_{ds} = \mu C_{ox} \frac{W}{L} V_t^2 (m-1) e^{(V_{gs} - V_T)/mV_t} (1 - e^{-V_{ds}/V_t}) \tag{2.44}$$

Note that the final term reduces to unity for $V_{ds} > 3V_t$.

Section 2.4 Conclusions

In this chapter, relevant fundamental aspects of semiconductor devices have been described. The operation of the MOS capacitor was reviewed, together with associated C-V measurement results and the extraction of some important parameters. The principles of the MOS transistor were also described in this chapter.

References

[1] A. Bar-Lev, *Semiconductors and Electronic Devices*. 3rd ed., Prentice Hall, 1993.

[2] S. M. Sze, *Semiconductor Devices: Physics and Technology*. 2nd ed., John Wiley & Sons, 2001.

[3] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge University Press, 1998.

[4] D. K. Schroder, *Semiconductor Material and Device Characterization*. John Wiley & Sons, 1998.

[5] M. Zerbst, "Relaxation effects at semiconductor-insulator interfaces," *Z. Angew. Phys.*, vol. 22, pp. 30–33, 1966.

[6] R. M. Swanson and J. D. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," *IEEE J. Solid-State Circuits*, vol. SC-7, pp. 146–153, 1972.

CHAPTER 3 CHARGE COUPLED SYNAPSE

Section 3.1 Introduction

A synapse is an electrochemical contact between neurons typically consisting of soma, axon, and dendritic trees as shown in Fig. 3.1. It constitutes the core processing unit within neuron cells. According to the synaptic weight, the spike emitted by the pre-synaptic neuron *i* is modulated to produce a time decaying output signal, commonly known as a post-synaptic potential (PSP) which will be transmitted to the post-synaptic neuron *j*. Due to the loading effect associated with the post-synaptic membrane, the rise and fall time constants of PSP are significantly different. A refractory period where another spike cannot be produced is present after the pre-synaptic neuron fires. Therefore, synaptic plasticity that modulates the impact of a PSP on the post-synaptic input, together with PSP associated time constants that closely mimics the charging/discharging processes associated with real spiking neurons, are essential learning characteristics and are taken into account in the proposed hardware implementation.

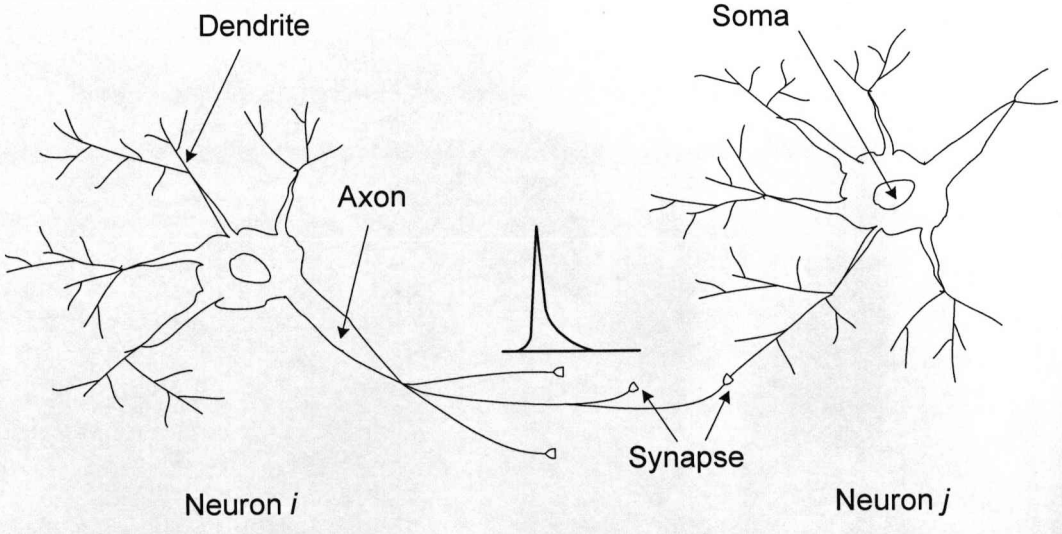


Fig. 3.1 Schematic drawing of a fragment of biological neural networks. The neuron *i* spikes and goes into a refractory state.

In this chapter a charge coupled synapse for spiking neural networks application is presented. In this endeavor, there are two major issues to be addressed: the synaptic device structure and the spiking behavior of the charge coupled synapse. The proposed silicon synapse is based on a two-phase charge transfer device with associated localized memory capability. The weighting functionality can be integrated into the first stage by means of a floating gate. A pre-synaptic spike to the second phase allows the charge under the first gate to drift onto the output terminal, which constitutes the interface with the point neuron, to produce a current or voltage spike.

The correspondence between the fundamental semiconductor processes and the required biological functionality is illustrated by theory in this chapter. The basic principles of the charge coupled synapse are demonstrated to exhibit a spiking characteristic of biological synapses. Issues related to the analytical modeling of the charge transfer and associated spike generation are investigated by using the Silvaco software package. The device simulations prove that the proposed synaptic device is able to capture the intrinsic dynamics of the biological synapses in spiking neural networks.

The rest of the chapter is organized as follows. In Section 3.2, the basic structure and design principles of charge coupled synapse are presented. Section 3.3 describes the voltage-dependent implementation of weight generation of the synapse. The charge transfer operation of the charge coupled synapse which mimics the weighting functionality of biological synapses is analyzed and modeled in Section 3.4. Section 3.5 deals with a description of the floating diffusion output stage of the synapse. In Section 3.6, the simulation results of the synaptic device are analyzed with the aid of the theory presented in the previous sections, and the biological behaviors are identified. Discussion and conclusions are given in Section 3.7.

Section 3.2 Charge Coupled Synapse Model

The proposed charge coupled synapse shown in Fig. 3.2 comprises a floating gate integrated into a two-stage charge transfer device with a heavily doped output

terminal. The two MOS capacitors (C1 and C2) are in close proximity, the first having charge storage capability and the second serving to 'clock' its reading. The synaptic weight is represented by the stored charge with associated voltage V_{ji} on the floating gate of the first capacitor where the charge would be added or removed through a tunneling process using well-established, read-only memory cell technology. The synaptic weight can therefore be updated by engineering the correlation between the pre-synaptic and the post-synaptic signals. The MOS capacitor C2 is controlled by the pre-synaptic spike V_i which releases the weight charge packet Q_W causing it to be dumped onto the output terminal. The device described here is an *n*-channel (excitatory synapse) but in practice a *p*-channel (inhibitory synapse) would be required since the weight charge stored is likely to be electrons on the floating gate electrode so the charge packet will be formed of holes. For ease of description we consider the charge slug to be formed by electrons and the models presented here are equally valid for *p*-channel operation with appropriate change of signs.

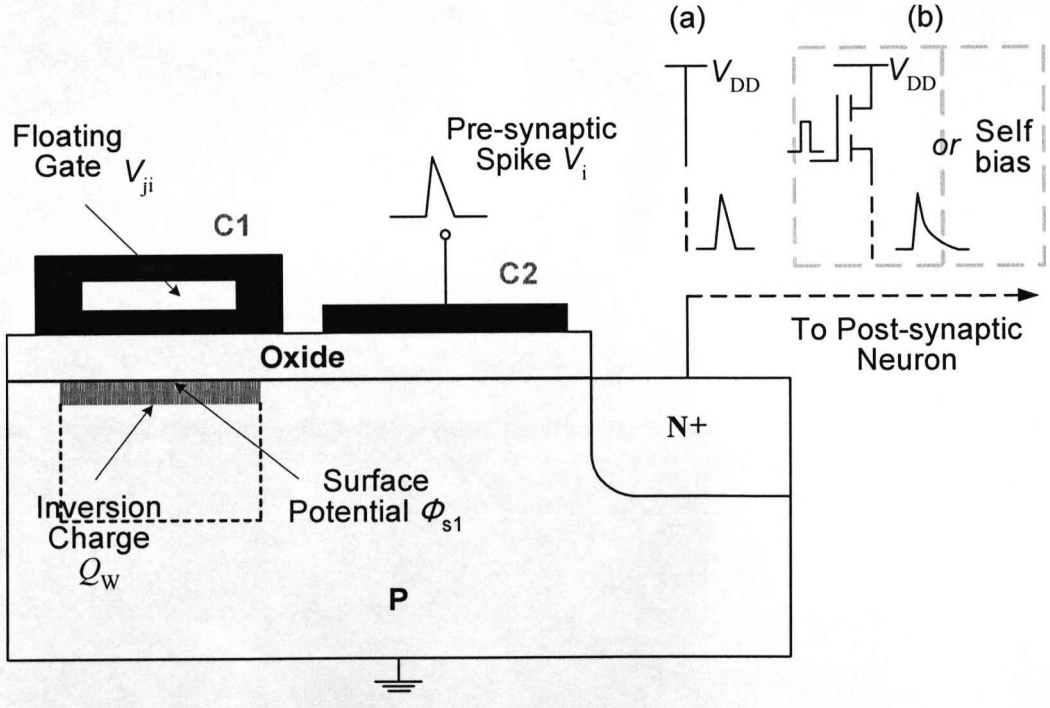


Fig. 3.2 Excitatory charge coupled synaptic model comprising two MOS capacitors (C1 and C2). An input pre-synaptic spike causes the weighted inversion charge to be dumped onto the output terminal. Output signal for different cases: (a) electron current spike for a positive power supply; (b) post-synaptic potential for a floating diffusion output node.

Charge detection occurs at the output terminal. This could be achieved by a positive bias or the gate of an MOS transistor such that the output terminal 'floats'. In conventional 'floating diffusion' output scheme of CCD technology, the dc bias is applied via an MOS transistor connected to the supply rail; that is, the transistor is switched synchronously with the C2, to apply and then remove the bias. A floating node can also be realized by connecting the output directly to the summing circuit. Referring to Fig. 3.2, in the first case, (a), the output electron current induced by the arrival of the (negative) charge is depicted. In the second case the bias is applied to output terminal via an MOS transistor, which is then switched off. The output will fall as the (negative) charge arrives but will relax over a much longer time duration because the charge can only be removed through the leakage current of the reverse-biased diode formed by the output terminal and substrate. An alternative approach that precludes the need for this bias circuitry is developed in Section 3.5. This approach relies on the capacitor divider effect to provide the 'self-bias' at the output terminal.

Section 3.3 Voltage-Dependent Synaptic Weight

In the first MOS capacitor C1, the weight charge packet Q_W stored at the oxide-semiconductor interface directly reflects the positive weight voltage V_{ji} associated with the charge stored in the floating gate. In order to illustrate the principle of the proposed structure in detail, a simplified structure (floating gate is omitted) shown in Fig. 3.3 is considered. It is assumed that the gate of the first MOS capacitor C1 is fixed at the weight voltage V_{ji}.

As described in the standard MOS theory [1], when a small V_{ji} ($V_{ji}>V_{FB}$) is applied, the MOS capacitor works in the depletion state. The mobile holes are repelled from the interface into the substrate, and the fixed ionized acceptors are exposed. With the increase the voltage V_{ji}, the oxide field begins to collect thermally generated electrons under the gate and the intrinsic surface begins to change into an *n*-type inversion layer.

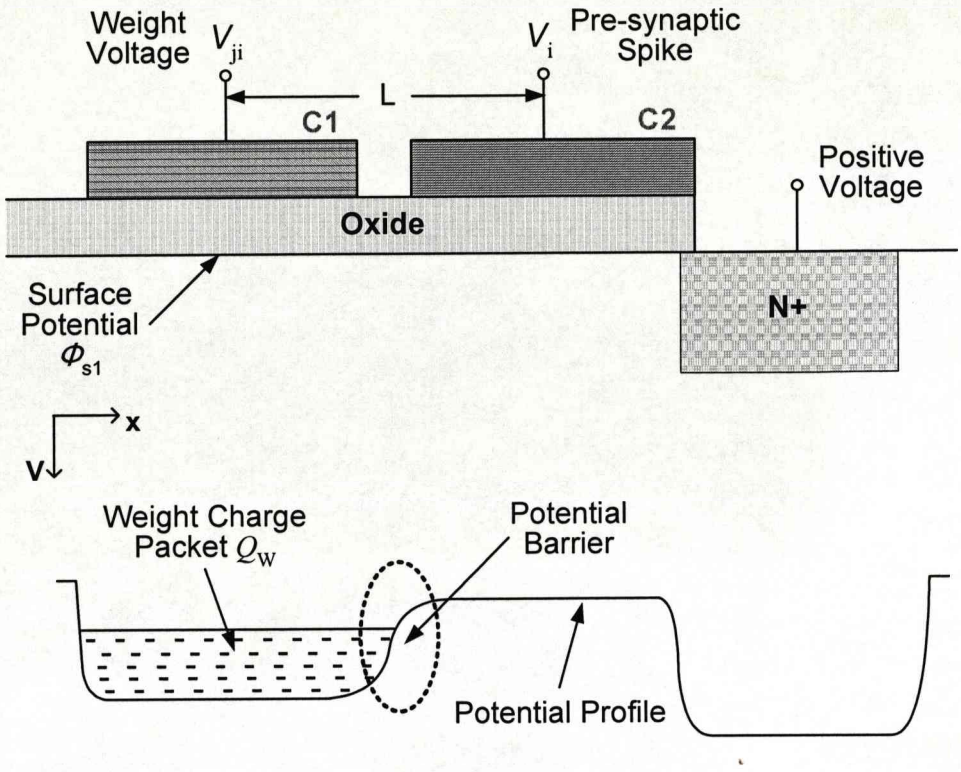


Fig. 3.3 A physical illustration of the weight charge packet Q_W in the charge coupled synapse (floating gate is omitted), in terms of the surface potential controlled by the weight voltage V_{ji}; The weight charge packet is stored locally due to the presence of the potential barrier between the MOS capacitors C1 and C2.

When the voltage V_{ji} is greater than the threshold voltage V_T:

$$V_T = V_{FB} + 2\phi_B + \frac{\sqrt{4\varepsilon_{si}\varepsilon_0 q\phi_B N_a}}{C_{ox}} \tag{3.1}$$

where V_{FB} is the flat-band voltage; ϕ_B is the bulk potential; N_a is the doping density of the substrate; C_{ox} is the oxide capacitance per unit area; q is the electronic charge, ε_{si} and ε_0 are respectively the semiconductor dielectric constant and vacuum permittivity, the MOS capacitor can operate in weak or strong inversion. The negative charge comprises of ionized acceptor atoms in the depletion region and free electrons which come from the thermal generation of electron-hole pairs in the depletion region. Note that the thermal generation process is rather slow especially for a well-treated semiconductor material. In strong inversion, the weight voltage V_{ji} enables a linear change of the inversion layer charge packet Q_W with charge stored on the floating gate. The MOS capacitor achieves the strong inversion state when the electron

concentration at the surface is much higher than the acceptor concentration. By ignoring the depletion charge, the inversion charge per unit area can be written as:

$$Q_W = (V_{ji} - V_T) \times C_{ox} \tag{3.2}$$

Alternatively, from Gauss's theorem and Poisson's equation the relationship between the weight charge density Q_W and the surface potential at the silicon-oxide interface ϕ_{s1} is obtained from (2.12) as:

$$Q_W = Q_0 \exp\left(\frac{\phi_{s1}}{2V_t}\right) \tag{3.3}$$

where $Q_0 = \sqrt{\frac{2q\varepsilon_{si}\varepsilon_0 V_t n_i^2}{N_a}}$; V_t is the thermal voltage (~0.026V at 300K); n_i is intrinsic carrier concentration and other symbols have their usual meaning.

Section 3.4 Synaptic Weighting Process

Consider the case with a charge packet Q_W underneath the first floating gate capacitor C1 where the surface is strongly inverted, as shown in Fig. 3.3. The pre-synaptic spike controls the gate of the second MOS capacitor C2. A fast rise time pre-synaptic signal V_i then arrives at the second input MOS capacitor driving it into the deep depletion state.

Now it is necessary to develop a relationship between the pre-synaptic spike V_i and the surface potential ϕ_{s2} of the MOS capacitor C2. Following the approach of Section 2.2.1, the relationship for the gate voltage V_i and the surface potential ϕ_{s2} can be expressed as:

$$V_i = V_{FB} + \phi_{s2} + \frac{\sqrt{2\varepsilon_{si}\varepsilon_0 q N_a \phi_{s2}}}{C_{ox}} \tag{3.4}$$

If the electrodes of MOS capacitors C1 and C2 are sufficiently close together such that their depletion regions interact, the weight charge packet Q_W can be transferred

from C1 to C2 due to the larger surface potential of C2 compared to C1: $\phi_{s2} > \phi_{s1}$. Therefore rewriting (3.4) yields:

$$V_i = V_{FB} + \phi_{s2} + \frac{\sqrt{2\varepsilon_{si}\varepsilon_0 qN_a\phi_{s2}}}{C_{ox}} + \frac{Q_W}{C_{ox}} \tag{3.5}$$

By solving (3.5), an expression for the surface potential ϕ_{s2} in terms of the applied voltage V_i and weight charge Q_W is obtained:

$$\phi_{s2} = V_1 + V_2 - \left[V_2^2 + 2V_1V_2\right]^{1/2} \tag{3.6}$$

where $V_1 = V_i - V_{FB} - \frac{Q_W}{C_{ox}}$ and $V_2 = \frac{qN_a\varepsilon_{si}\varepsilon_0}{C_{ox}^2}$.

The deeper surface potential well formed in the silicon under the second gate, causes the charge packet Q_W to start transferring from the first capacitor to the second and subsequently to the output terminal, which controls the membrane potential of the post-synaptic neuron. During this transfer process, the lateral charge profile in the charge coupled synapse will not be uniform as demonstrated in Fig. 3.4.

The charge motion takes place due to three physical mechanisms, namely self-induced drift, fringing field drift, and thermal diffusion [2-4]. Two sources of the electric field act on the charge packet: the self-induced field ξ_S due to the distribution of the mobile charge itself and the fringing field ξ_F due to the externally applied potentials on the electrodes. The treatment here follows closely that of [2-4]. Therefore, the current density in the charge coupled synapse can be written as three terms:

$$J = -\mu n_W(x,t)\xi_S(x,t) - \mu n_W(x,t)\xi_F(x,t) - D\frac{\partial n_W(x,t)}{\partial x} \tag{3.7}$$

where $n_W(x,t)$ is the weight charge density per unit area in the storage well; x is the direction of charge propagation along the interface; D is the diffusion coefficient appropriate to the surface; μ is the surface mobility.

The transport equation will be based on the divergence equation for current:

$$\frac{\partial n_W(x,t)}{\partial t} = -\nabla \bullet J \tag{3.8}$$

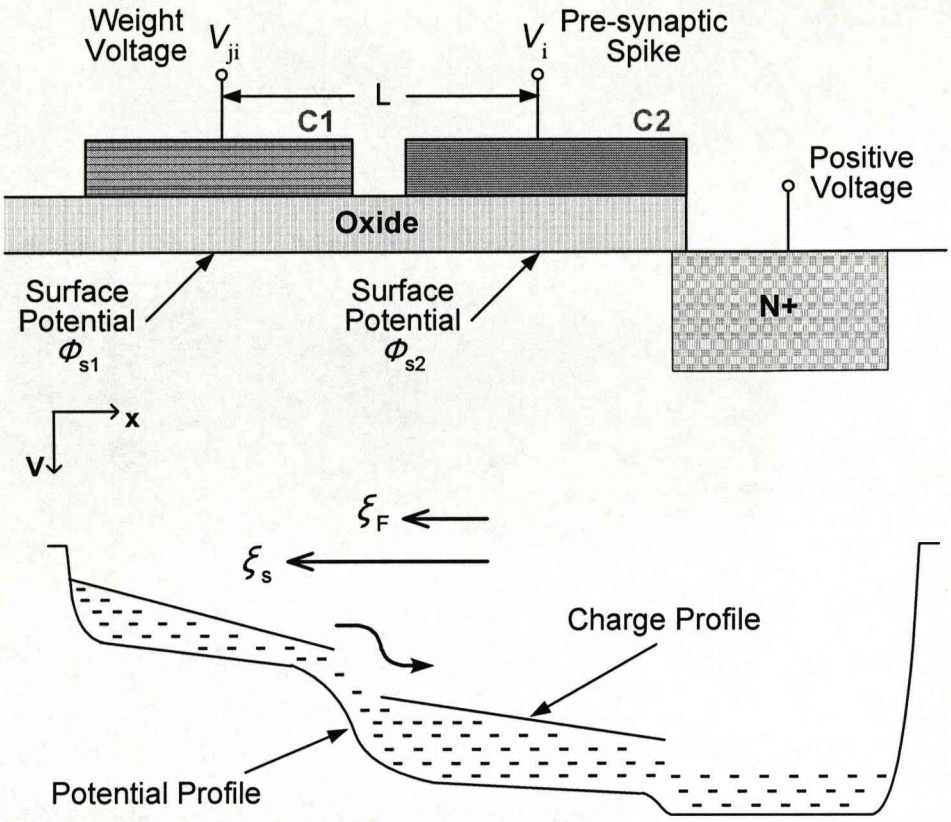


Fig. 3.4 A physical illustration of the synaptic weight transfer process in terms of the variations of the surface potential controlled by the pre-synaptic spike V_i. The weight charge packet seeks deeper potential well and is transferred toward the output terminal when a pre-synaptic spike is applied.

Combining (3.7) and (3.8) gives the basic continuity equation for the decay of charge in the charge coupled synapse:

$$\frac{\partial n_W(x,t)}{\partial t} = \mu \frac{\partial}{\partial x}\left[n_W(x,t)\xi_S(x,t)\right] + \mu \frac{\partial}{\partial x}\left[n_W(x,t)\xi_F(x,t)\right] + D\frac{\partial^2 n_W(x,t)}{\partial x^2} \qquad (3.9)$$

By using the Einstein relationship between diffusion and mobility:

$$D = \mu \frac{kT}{q} \qquad (3.10)$$

where k, T and q are the Boltzmann's constant, absolute temperature, and electronic charge. Then, one reaches this equation:

$$\frac{\partial n_W(x,t)}{\partial t} = \mu \frac{\partial}{\partial x}[n_W(x,t)\xi_S(x,t)] + \mu \frac{\partial}{\partial x}[n_W(x,t)\xi_F(x,t)] + \mu \frac{kT}{q} \frac{\partial^2 n_W(x,t)}{\partial x^2} \quad (3.11)$$

In order to get an approximated analytical solution to the above equation, it is convenient to consider the three transfer mechanisms separately.

Section 3.4.1 Self-induced Drift

The weight charge packet Q_W leaving the right hand side of the MOS capacitor C1 will cause an associated local increase in surface potential such that the potential will vary with lateral distance x. Essentially the charge transport itself induces a lateral electric field. The potential variation associated with the self-induced field, which dominates the transfer process, is shown in Fig. 3.4.

By taking the gradient along the interface of the surface potential ϕ_{s1} of C1, the magnitude of the self-induced field ξ_S can be obtained since ϕ_{s1} is a function of the weight charge density. The surface potential under the first gate can be written as:

$$\phi_{s1}(x,t) = \phi_{s1,0} - \frac{q}{C_{ox}} n_W(x,t) \quad (3.12)$$

where $\phi_{s1,0}$ is the surface potential with no electron at the interface, given by:

$$\phi_{s1,0} = \frac{qN_a W_{d1}^2}{2\varepsilon_{si}\varepsilon_0} \text{ or } \phi_{s1,0} = 2\phi_B = 2V_t \ln\frac{N_a}{n_i} \quad (3.13)$$

where W_{d1} is the depletion layer width of MOS capacitor C1. Note that the depletion capacitance C_{d1} has been neglected since it is much bigger than the oxide capacitance C_{ox}.

The self-induced electric field ξ_S under the first gate is then:

$$\xi_S(x,t) = -\frac{\partial \phi_{s1}(x,t)}{\partial x} = \frac{q}{C_{ox}} \frac{\partial n_W(x,t)}{\partial x} \quad (3.14)$$

Thus, the current density per unit length due to the self-induced field is:

$$J = -\mu n_W(x,t)\frac{q}{C_{ox}}\frac{\partial n_W(x,t)}{\partial x} \tag{3.15}$$

The transport equation is given by:

$$\frac{\partial n_W(x,t)}{\partial t} = \mu\frac{q}{C_{ox}}\frac{\partial}{\partial x}\left[n_W(x,t)\frac{\partial n_W(x,t)}{\partial x}\right] \tag{3.16}$$

This is in the form of diffusion equation:

$$\frac{\partial n_W(x,t)}{\partial t} = \frac{\partial}{\partial x}\left[D_{eff}\frac{\partial n_W(x,t)}{\partial x}\right] \tag{3.17}$$

where $D_{\text{eff}} = \mu q n_{\text{W}}(x,t)/C_{\text{ox}}$ is a function of charge concentration. Therefore, the charge transfer process due to the self-induced electric field can be treated as a simple diffusion-like process.

Comparing the coefficient D_{eff} with thermal diffusion constant D:

$$\frac{D_{eff}}{D} = \frac{q^2 n_W(x,t)}{kTC_{ox}} \tag{3.18}$$

It can be concluded that for a filled potential well, D_{eff} is much larger than D, and the charge transfer process due to the self-induced field is much faster than thermal diffusion [5]. An expression for the effective transit time is obtained by the analytical solution of (3.11):

$$\tau_{SID} = \frac{L^2 C_{ox}}{\mu Q_W}\left(\frac{Q_W}{Q_f} - 1\right) \tag{3.19}$$

where L is the distance between the centres of the two electrodes (see Fig. 3.4); μ is the average carrier mobility; Q_{f} is the final charge remaining in the storage well. Therefore, the total charge remaining under the first gate will decay according to the following expression:

$$Q_f(t) = \frac{Q_W L^2 C_{ox}}{L^2 C_{ox} + \mu Q_W \tau_{SID}} \tag{3.20}$$

Section 3.4.2 Thermal Diffusion

Once the self-induced drift coefficient D_{eff} is smaller than the thermal diffusion constant D, the level of charge in the storage well (under the first gate) is less than the average concentration due to the dopant; that is when:

$$Q_f < \frac{DC_{ox}}{\mu} \tag{3.21}$$

For this condition, the charge transfer process becomes to be dominated by the comparatively slow thermal diffusion aided by the fringing field drift. In our case, this corresponds to a low concentration of charge which is essentially outside the operational dynamic range.

The charge transfer due to the thermal diffusion alone can be determined by means of Fourier analysis of the charge profile. The analysis of the thermal diffusion process shows that for an initially uniform electron concentration $n_{W,0}$, the profile $n_W(x,t)$ remaining under one electrode has the asymptotical expression in time:

$$n_W(x,t) = \frac{4n_{W,0}}{\pi}\cos\left(\frac{\pi}{2L}x\right)\exp\left(-\frac{\pi^2 D}{4L^2}t\right) \tag{3.22}$$

The total number of electrons remaining at time t is:

$$n_W(t) = \frac{8}{\pi^2}n_{W,0}\exp\left(-\frac{\pi^2 D}{4L^2}t\right) \tag{3.23}$$

Aside from a small amount of charge that decays very quickly, the decrease of the total charge due to thermal diffusion is exponential with time constant:

$$\tau_{TD} = \frac{4L^2}{\pi^2 \mu V_t} \tag{3.24}$$

As shown in Fig. 3.5, the charge transfer process finishes up with the loss of a small amount of charge and the re-establishment of the potential barrier. Note that Fig. 3.5 still represents a non-equilibrium condition. The overall relaxation of the system is a

result of thermal generation of electron-hole pairs under the electrodes. This is considered in Chapter 4.

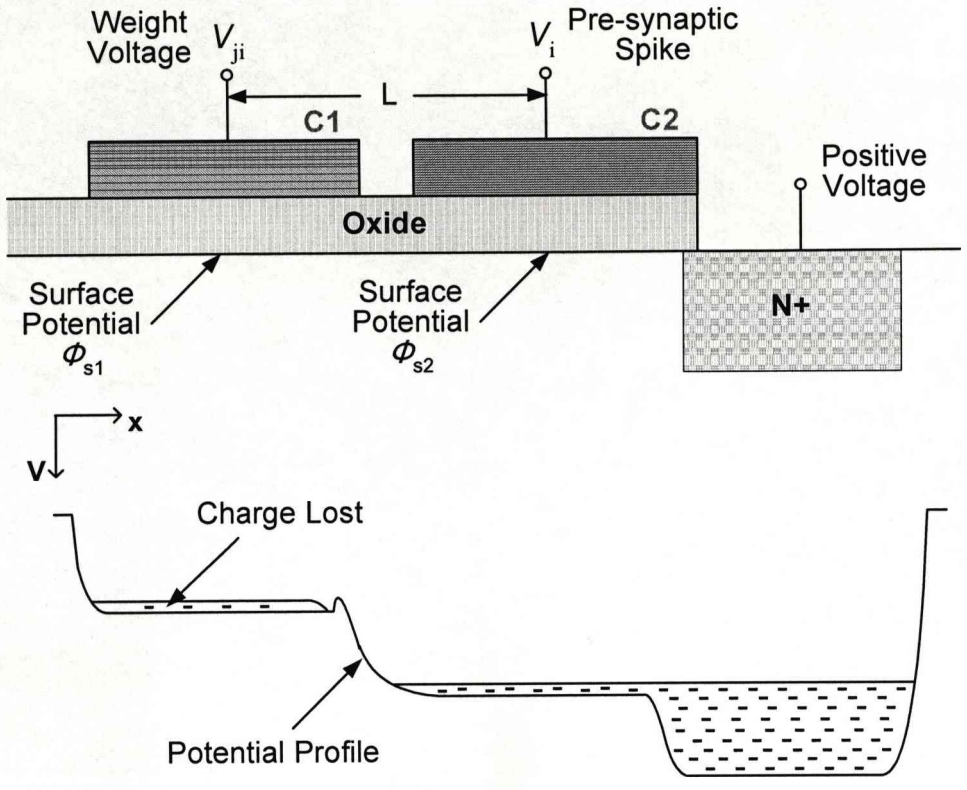


Fig. 3.5 A physical illustration of the charge profile at the end of the charge transfer process. A fragment of the charge packet is lost in the storage well of the MOS capacitor C1.

Section 3.4.3 Fringing Field Drift

Due to the coupling of the electrostatic potential caused by the interaction between adjacent electrodes, the fringing field is present during the charge transfer process even when the charge concentration is low, and has the same direction as self-induced electric field. A simple estimate of the fringing field is given by Carnes *et. al.* [2]. The magnitude of this fringing field is determined by both an approximate analytical technique and computer simulation. The empirical formula which permits rapid calculation of the minimum fringing field is:

$$\xi_F = 6.5\frac{C_{ox}\Delta V}{L^2}\left[\frac{5W_{d2}/L}{5W_{d2}/L+1}\right]^4 \tag{3.25}$$

where ΔV is the voltage difference between the adjacent electrodes; W_{d2} is the width of the depletion layer in the centre of the MOS capacitor C2. Note that the fringing field is larger for thicker oxide, smaller electrode length and greater clock swing. For larger electrodes (several microns), useful fringing fields require rather thick gate oxide, typically 50nm-100nm, but considerably thinner oxides are appropriate for scaled devices. In a certain range of electrode lengths and substrate doping levels, the single carrier transit time, which is the time a single carrier would require to drift from one end of the electrode to the other under the influence of a fringing field, is small enough to indicate that fringing field drift may greatly aid the transfer process.

The transfer of charge from one potential well to another is the most important factor affecting both the speed of operation and the efficiency attainable in the charge coupled synapse. In the absence of a fringing field, the self-induced drift dominates the transfer of most of the charge, and thermal diffusion is responsible for the transfer of the last few percent of charge [6]. By considering just the self-induced drift process, one can conclude that reasonable transfer efficiencies could only be achieved by transferring just a portion of the charge [7].

Section 3.5 Floating Diffusion Output Stage

Since charge in the inversion layer of C1 can only be established through thermal generation of electron-hole pairs in the associated depletion layer, the lateral drift of weight charge packet Q_W will result in a transient current/voltage (or spike) at the output, as the density of charge in the inversion layer diminishes with time. Hence, the output of an unloaded synapse will exhibit a spike shape characteristic whose magnitude is affected by the density of charge in the inversion layer, and consequently the magnitude of V_{ji} through the relationship $Q_W=C_{ox}(V_{ji}-V_T)$. This correlation between V_{ji} and Q_W can be used to mimic synaptic plasticity and results will be presented later in support of this assertion. Furthermore, immediately after the

charge has transferred, the charge coupled synapse is in a state of non-equilibrium with deep depletion conditions under the gate of C1 and C2 and the output node. The overall relaxation to equilibrium can be thought of as mimicking a refractory period and this is dominated by the slowest time constant which is associated with the deep depletion conditions on the MOS capacitors C1 and C2.

Consider the case where the output signal from the charge coupled synapse is detected at the floating diffusion structure as follows. Referring to Fig. 3.6, the capacitor divider effect of the series combination of the overlap capacitance, C_{ov} and the floating diffusion, C_{FN}, is exploited such that a positive potential is induced on the floating diffusion node (FDN), by the fast rise time pulse at the second electrode. Note that the output/substrate junction is driven into reverse bias and this is also a non-equilibrium dynamic condition. The capacitance between the FDN and the substrate, C_{FN}, is dependent on the length of FDN region L_{FN}. The coupling capacitance, C_{ov}, is a function of the length of overlapping between the FDN and the second gate L_{ov}. The capacitance C_{FN} can be reduced, by reducing the FDN length L_{FN} to the minimum feature size, to provide the necessary 'self-bias' at the output. This approach precludes the need for bias circuitry and maintains the highly compact nature of the synapse with a commensurate reduction in power consumption.

Therefore the pre-synaptic spike V_i to the second gate causes a significant increase of the potential, in tandem with the surface potential induced in the second MOS capacitor C2. As soon as the deeper potential well is formed under the second gate, the weight charge packet Q_W is transferred onto the FDN region, where it is converted into a voltage [8]. Therefore the potential on the FDN is decreased. Dumping the charge Q_W onto the FDN with capacitance C_{FN} produces an associated voltage change:

$$\Delta V = \eta \frac{Q_W}{C_{FN}} \tag{3.26}$$

where η is the transfer efficiency.

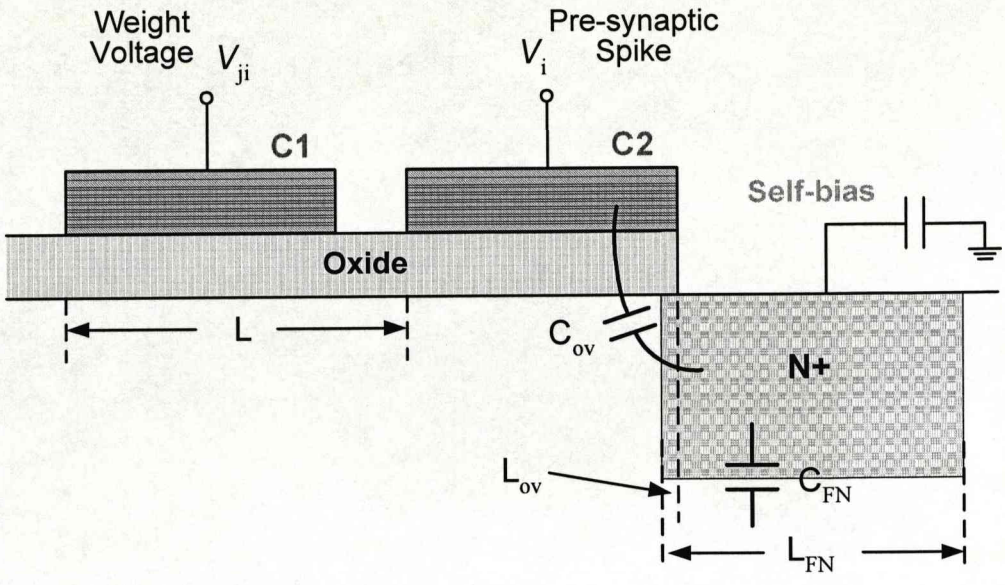


Fig. 3.6 Schematic diagram of the charge coupled synapse, showing the equivalent circuit of the floating diffusion output stage. The output is connected to the gate of a transistor, represented by a capacitor.

The collection of charge relies on the natural capacitive coupling from the gate of C2 to the FDN. For a voltage on the second gate, V_i, the voltage coupled onto the FDN can be estimated as:

$$V_{FN} = V_i \frac{C_{ov}}{C_{ov} + C_{FN}(V_{FN})} \tag{3.27}$$

where

$$C_{ov} = \frac{\varepsilon_{ox}\varepsilon_0}{t_{ox}} W L_{ov} \tag{3.28}$$

and

$$C_{FN} = \sqrt{\frac{qN_a\varepsilon_{si}\varepsilon_0}{2(V_{bi} + V_{FN})}} W L_{FN} \tag{3.29}$$

Equation (3.27) constitutes a quadratic equation which is complicated to solve due to the presence of the built-in voltage, V_{bi}. It is easiest to substitute a value for V_{FN} and hence obtain a corresponding value for V_i; V_{FN} can then be adjusted until the desired value of V_i is obtained.

The transient voltage appearing on the FDN will then be:

$$V_o(t) = V_{FN}(t) - \Delta V(t) \tag{3.30}$$

where $\Delta V(t)$ is given by (3.26).

The form of the transient voltage on the FDN is illustrated schematically in Fig. 3.7. The fast rise time is dictated by the dielectric relaxation time for majority carriers in the FDN and substrate. The falling edge of the spike is dictated by two processes: the arrival of the electron charge packet with a short time duration τ_{f1}, and the subsequent relaxation of the voltage by leakage in the diode formed by the FDN and the substrate with time duration τ_{f2}.

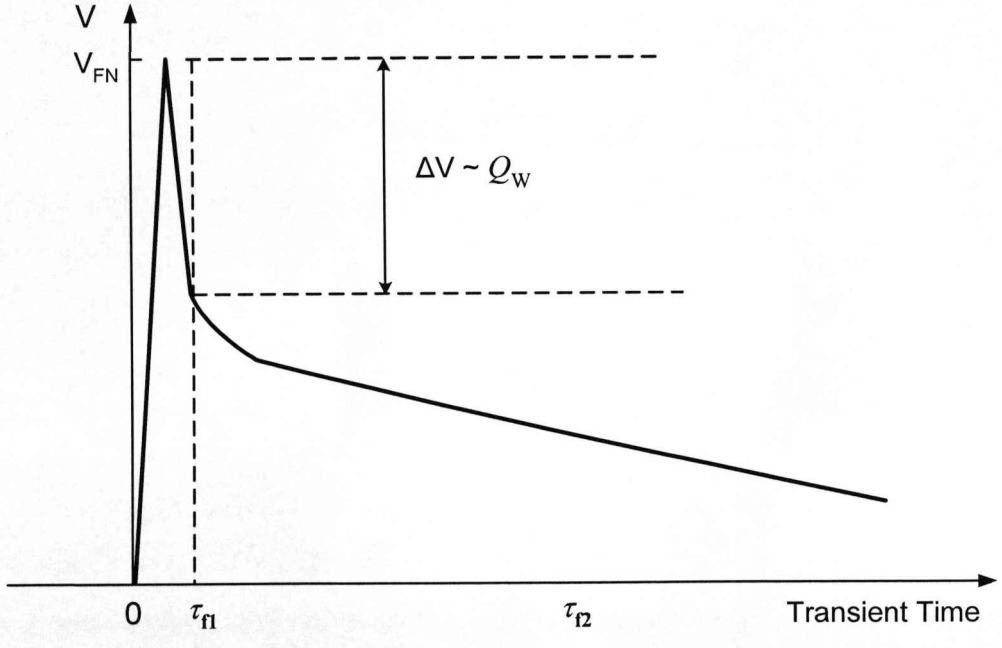


Fig. 3.7 Schematic illustration of the FDN potential.

By considering the drift due to the self-induced field [4], the characteristic time τ_{f1} for the transit through the charge coupled synapse is estimated by (3.19). The relaxation time τ_{f2} can be found by considering the leakage of charge from the one-sided abrupt FDN/substrate diode due to the thermal generation process:

$$I_{leak} = -C_d(V_o)\frac{dV_o}{dt} \tag{3.31}$$

where $$C_d(V_o) = \frac{\varepsilon_{si}\varepsilon_0}{W_d(V_o)} \tag{3.32}$$

The leakage current I_{leak} can also be expressed as:

$$I_{leak} \sim \frac{qn_i W_d(V_o)}{\tau_g} \tag{3.33}$$

where $W_d(V_o)$ is the non-equilibrium, reverse bias depletion width, τ_g is the generation lifetime, and n_i is the intrinsic carrier concentration.

Substituting (3.32) and (3.33) into (3.31) and separating variables yields:

$$\int_0^{\tau_{f2}} dt = -\tau_g \frac{N_a}{2n_i} \int_{V_o}^{0} \frac{dV_o}{V_{bi} + V_o} \tag{3.34}$$

Thus the longer time τ_{f2} can be estimated as:

$$\tau_{f2} = \tau_g \frac{N_a}{2n_i} \ln\left(\frac{V_{bi} + V_o}{V_{bi}}\right) \tag{3.35}$$

The slow decay of the voltage spike implied by (3.35), closely models the transient characteristic or PSP associated with real neurons without requiring any additional RC networks. There are well-established 'lifetime quenching' engineering solutions (used in power device technology) to tailor τ_g to the required value to set the relaxation time constants to those expected for a realistic PSP. Usually these techniques are employed to reduce the recombination lifetime whereas here, the aim is to reduce the generation lifetime. The techniques rely on introducing deep levels into the band gap of Si. Such levels act as 'stepping' stones for generation/recombination (G/R) processes whereby charge carriers move between conduction and valence bands [10]. The G/R processes are controlled by the same charge balance statistics but there are subtle differences between the two limiting cases. In general, a deep level at mid-gap implies reciprocity of the two processes. In our case, a trap at mid-gap is preferred as this maximizes the generation process. Note that the total charge transfer time can be of the order 10ns which is at least an order of magnitude less than the decay period in PSP. Therefore an asymmetric output spike is maintained even for subthreshold operation. As described in the previous section, the charge coupled synapse requires a

low doped substrate to produce significant fringing fields between the electrodes for efficient charge packet transfer. The time τ_{f2} is therefore likely to be prohibitively long without lifetime quenching. PSP duration of milliseconds is achievable by this means. Simulation results from the analysis presented in this section are deferred to Section 3.6.4.

A multi-input transistor or neuMOS [11] has been considered as a summing and thresholding device – essentially a device that acts like a point neuron. Synapses can be assigned to each of the neuMOS inputs, in principle, with fan-in limited only by the physical size of the transistor. This latter point arises because, unlike other applications for such transistors, it does not require fast operation of the device as the biologically inspired circuit relies on parallelism rather than speed. The pre-synaptic signal can be fed through to induce a positive potential on the FDN: a self-induced virtual bias. The post-synaptic potential read by one of the sub-gates of the neuMOS can be used to trigger the summing and thresholding neuron cell. When the pre-synaptic spike V_i goes down to zero in milliseconds, the post-synaptic potential drops which could enable the neuron cell to depress. Then the potential resets to its equilibrium value slowly; alternatively a reset transistor would be employed to reset the FDN. The proposed charge coupled synapse with 'self-biasing' FDN is suitable for use with the neuMOS and this approach is feasible, but there are some challenges in implementing it due to the nature of the signal generated at the output which requires additional circuitry to de-embed the information that can then be assigned to the inputs of the neuMOS.

Section 3.6 Simulation Study and Results Analysis

The synaptic device model is set up by using the Devedit in Deckbuild interactive environment. Referring to Fig. 3.2, two *n*-polysilicon gates with 0.5μm spacing are placed on a 100nm thick oxide layer. The *p*-type substrate is doped at $N_a=10^{15}cm^{-3}$. The length of the output terminal, doped with density $N_d=10^{19}cm^{-3}$, is set to 3μm.

Section 3.6.1 Synaptic Weight Generation

As described in Section 3.3, a 'permanent' packet of charge Q_W utilized to produce synaptic plasticity is stored at the silicon-oxide interface of the first capacitor C1 and represents the positive weight voltage V_{ji}. Firstly the dynamic operating range for the weight charge density in the MOS capacitor C1 can be established by using (3.3). The minimum charge density, $n_{W,min}$~1.9×10^{10}cm^{-2}, is calculated by considering the strong inversion condition, ϕ_{s1}~$2\phi_B$ where $\phi_B=V_t\times\ln(N_a/n_i)$ is the bulk potential with value 0.29V in this case. Standard *p-n* junction theory gives a depletion width under this condition of W_{d1} of 0.87μm. Therefore the depletion charge per unit area is estimated as N_aW_{d1}~8.8×10^{10}cm^{-2}. It is worth noting that this is in excess of the parasitic interface charge associated with the surface states at the oxide-semiconductor interface which can be engineered to be of the order 10^{10}cm^{-2}. The upper value of weight charge is set by oxide leakage current considerations. The charge stored on the gate of C1 will establish an electric field, the maximum value of which will be ~6MV/cm which translates into a maximum level of charge of the order 10^{13}cm^{-2}. Higher field would result in self-induced Fowler-Nordheim leakage and corruption of the weight. Such leakage would arise via the thin tunneling oxide required on top of the floating gate. The dynamic range of the weight charge per unit area can therefore be estimated to be in the range 10^{11}cm^{-2} < n_W < 5×10^{12}cm^{-2}.

The calculations of the synaptic weight generation are performed in the ATLAS 2D device simulation framework (S-Pisces). Appropriate semiconductor physical models, such as field-dependent mobility and Shockley-Read-Hall generation/recombination are included. A positive voltage of 2.5V is applied to the N+ output terminal, and a voltage of 1V is applied to the gate of C1, representative of the synaptic weight V_{ji}. Fig. 3.8 schematically shows the electron concentration profile of the charge coupled synapse, which is in equilibrium before the pre-synaptic signal arrives. The induced weight charge packet Q_W is stored locally due to the potential barrier between two MOS capacitors C1 and C2. The magnitude of the weight charge concentration as a function of weight voltage V_{ji} on the gate of C1 is shown in Fig. 3.9. The threshold for the MOS capacitor C1 is 0.13V, as indicated clearly in the figure. For 1V weight voltage V_{ji}, the electron volume concentration is 9.47×10^{16}cm^{-3}, corresponding to the

density at the surface of $1.86\times10^{11}cm^{-2}$ ($=2.97\times10^{-8}C/cm^2$). This is coincident with the theory where the threshold voltage calculated by (3.1) is 0.13V and the density of charge obtained from (3.2) is $1.9\times10^{11}cm^{-2}$ by ignoring the depletion charge.

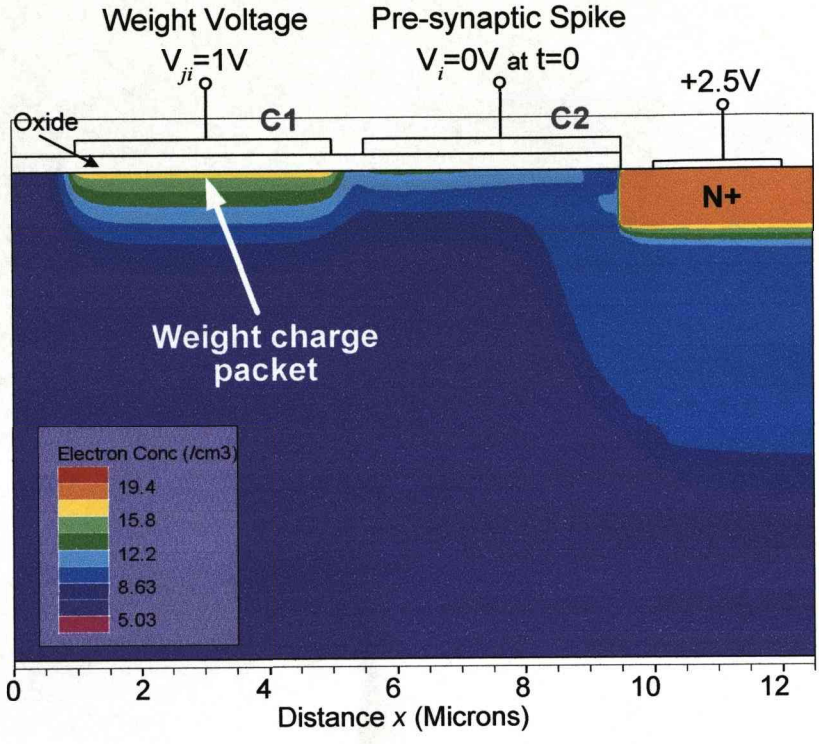


Fig. 3.8 Schematic drawing of the electron concentration profile before the pre-synaptic spike arrives at C2. The weight charge packet Q_W, induced by V_{ji}, are in the storage well of the MOS capacitor C1. A positive voltage of 2.5V is applied to the N+ output terminal.

Section 3.6.2 Synaptic Weighting Process

Firstly it is necessary to estimate the value for the voltage pulse on the gate of MOS capacitor C2 by recognizing that $\phi_{s1}\sim2\phi_B$ (~0.58V). By setting $\phi_{s2}=\phi_{s1}+0.1V$, where a difference in potential of $4V_t$ has been assumed, the lower limiting value of V_i for synaptic operation is found to be 1.6V, assuming no work function difference in (3.6). Work function difference will depend on the choice of gate material; *p*-type polySi would provide little difference.

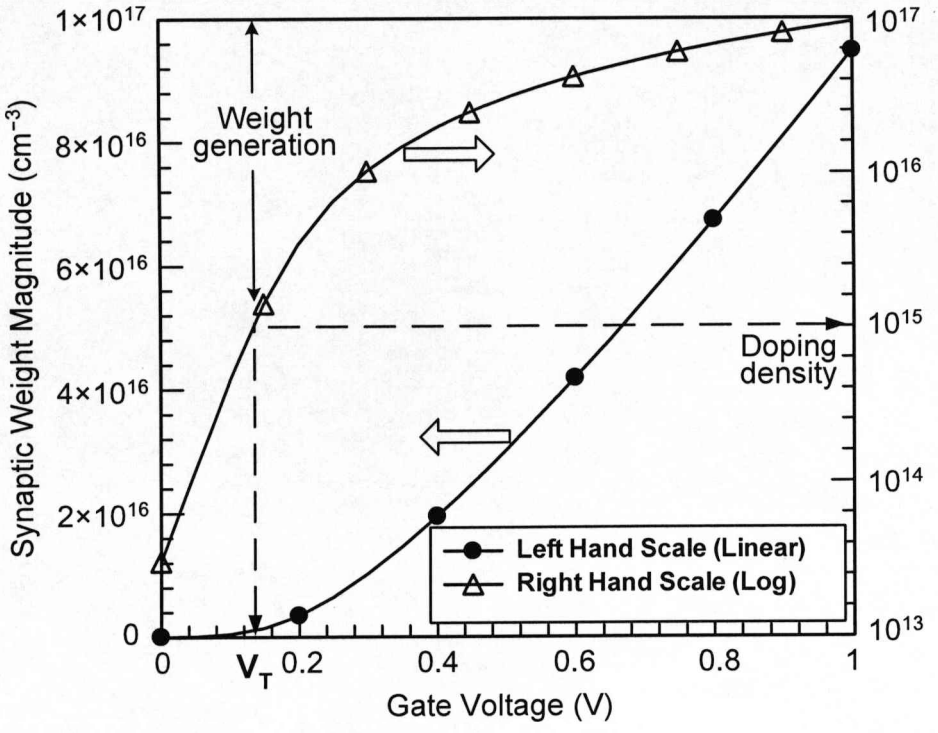


Fig. 3.9 Weight charge packet generation at the interface of the storage well when weight voltage V_{ji} is applied to the gate of the MOS capacitor C1.

It turns now to an estimate of the transit time of the weight charge Q_W. Substituting typical values of t_{ox}=100nm, μ=400cm^{2}V^{-1}s^{-1}, and L=1μm and Q_f =0.1Q_W in (3.19) gives times of the order of tens of picoseconds over a range of charge 10^{11}cm^{-2} < Q_W < 3×10^{12}cm^{-2}. A thinner gate oxide of 25nm still gives values less than 1.8ns. Such times are considerably faster than the time constants associated with relaxation due to thermal generation and therefore the assumption that the charge transfer process is essentially instantaneous compared to thermal processes is valid.

The gate oxide thickness t_{ox}, depletion region width W_{d2} and gate electrode length and spacing need to be designed to provide a fringing field of >~50kV/cm. Fig. 3.10 shows plots of (3.25), for oxide thicknesses of 100nm and 20nm, with a constant voltage on C2 of 2.5V and a depletion width obtained from ϕ_{s2}. The result indicates that operation with gate oxides of 20nm is feasible. Thus the same gate oxide

thickness can be used for both charge coupled synapses and accompanying MOS transistors, assuming a minimum feature size of 1μm.

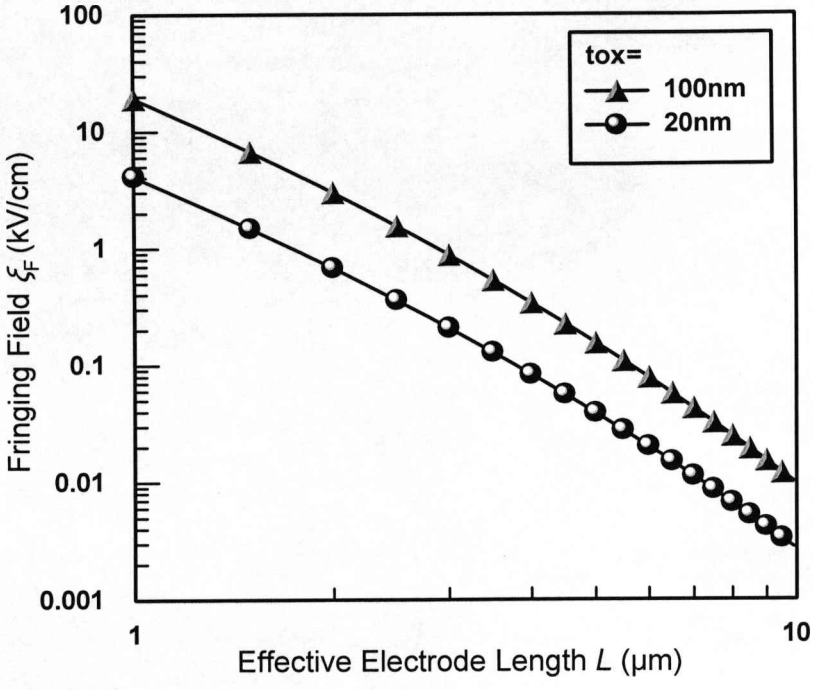


Fig. 3.10 Fringing field for two different oxide thicknesses (t_{ox}=100nm; t_{ox}=20nm).

(1) Pre-synaptic Spike with 1ns Rise Time

As soon as the pre-synaptic voltage spike V_i with the amplitude of 2.5V, shown in Fig. 3.11, arrives at the gate of the second MOS capacitor C2, the silicon beneath the gate of this capacitor is driven into deep depletion and a deeper potential well is formed as indicated in Fig. 3.12. The potential barrier between two MOS capacitors vanishes and the lateral potential profile is not uniform, as a result of the charge loss from the right hand side of the first capacitor C1: this brings the self-induced field to the device. Note that the self-induced field dominates the transfer of most of the charge, with the aid of fringing field. When the pre-synaptic spike reaches 2.5V, the potential under the second gate increases towards its maximum and 45% of the weight charge packet has drifted, as indicated in Fig. 3.13. At 2ns, 87.5% of the weight charge

packet has transferred. When the weight charge density drops to $5.6\times10^{9}cm^{-2}$ at 4ns, it can be observed that the thermal diffusion starts to dominate the transfer and this occurs towards the end of the charge transfer process. At 5ns, the potential under two MOS capacitors increases to its maximum and 99.6% of the weight charge packet has transferred. A transfer efficiency of 99.9% is achieved due to the thermal diffusion exponential with time constant 6ns. As schematically shown in Fig. 3.14, the majority of the weight charge Q_W has flowed out at the end of the charge transfer process, and the electron concentration in both MOS capacitors is about $10^{12}cm^{-3}$ that is much smaller than the acceptor density in the *p*-substrate. It should be noted that the charge transfer process is in 10ns after which the charge coupled synapse starts to recover. The potential curve in Fig. 3.12 for 1ms when the pre-synaptic spike drops to 0V clearly indicates that the depletion layer width of the MOS capacitor C1 has not fully recovered to equilibrium.

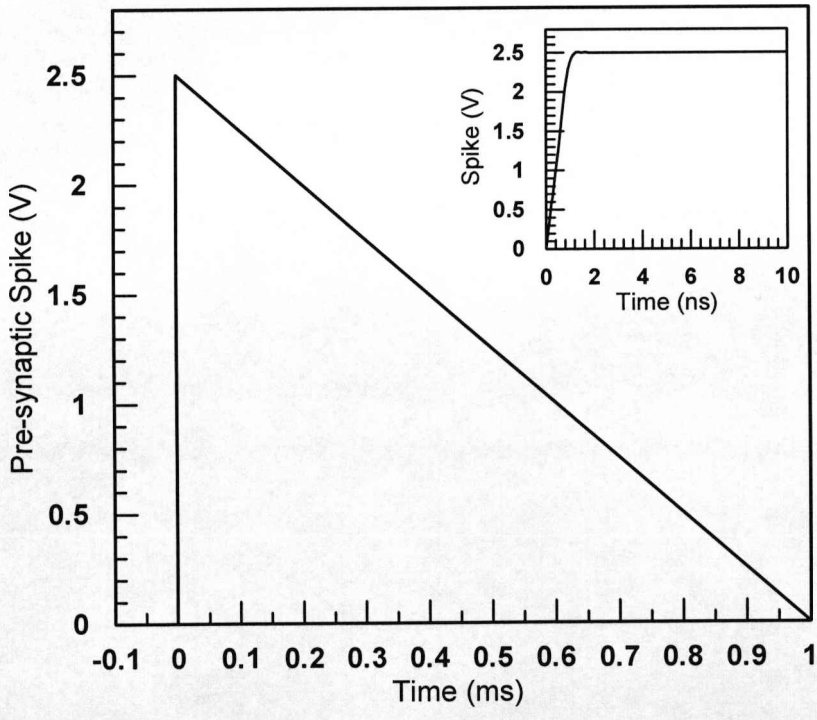


Fig. 3.11 Pre-synaptic spike V_i to the gate of the MOS capacitor C2. The spike is ramped from 0V to 2.5V over a period of 1ns, and lasts 9ns, as indicated in the inset. The fall time is 1ms.

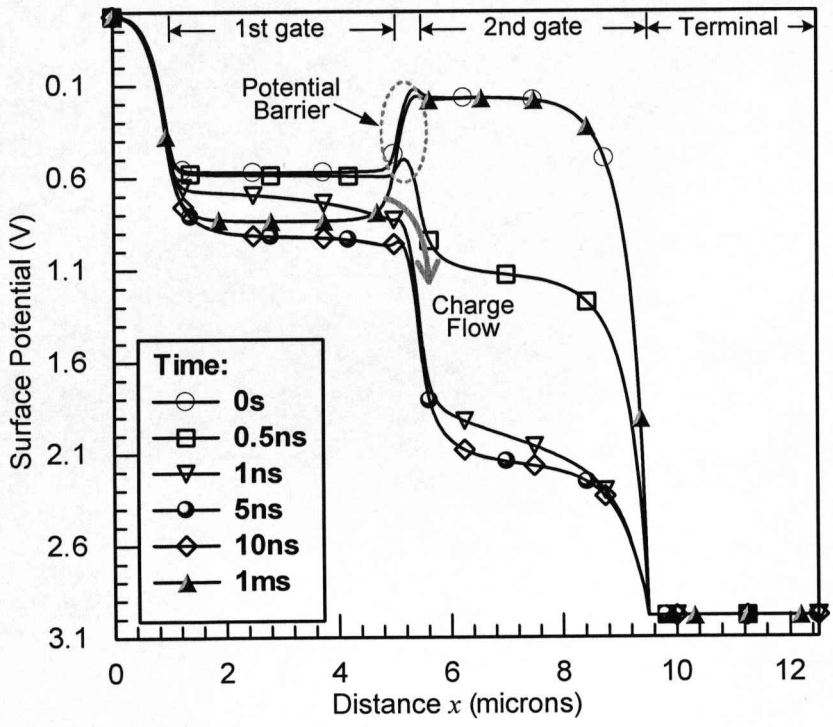


Fig. 3.12 Potential profiles along the surface of charge coupled synapse at different spike times.

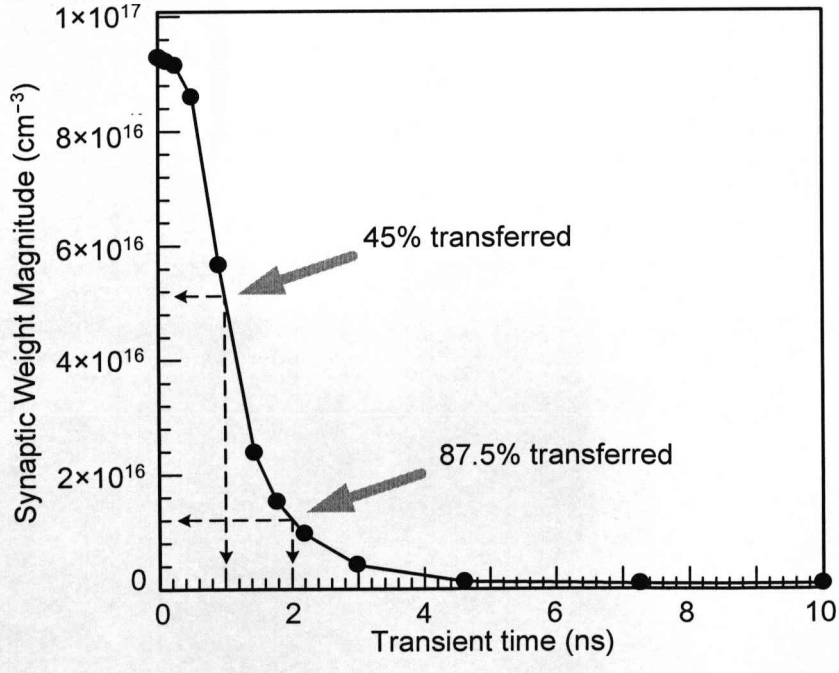


Fig. 3.13 Electron concentration under the gate of MOS capacitor C1 over spike time. The weight charge packet Q_W flows toward the output terminal when the pre-synaptic spike arrives at the gate of C2.

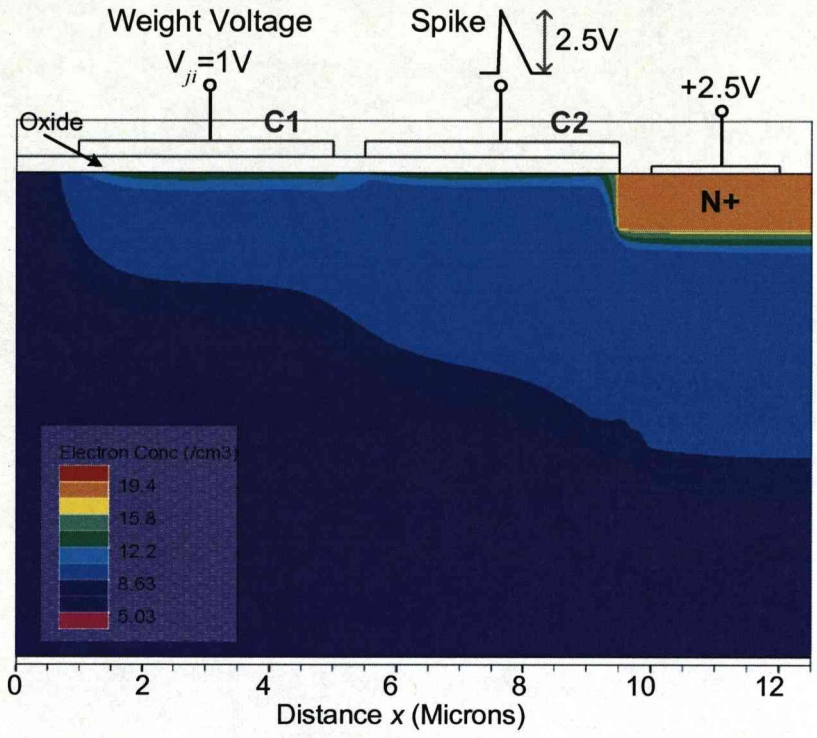


Fig. 3.14 Schematic drawing of the electron concentration in the silicon of the charge coupled synapse after the synaptic weighting process. The majority of weight charge packet Q_W has transferred.

(2) Pre-synaptic Spike with Slow Rise Time

In this simulation study, a relative slow pre-synaptic spike shown in Fig. 3.15 is applied to the gate of the MOS capacitor C2. The spike is ramped from 0V to 2.5V over a period of 1μs, and lasts 1μs. The fall time is 1ms, consistent with the typical time regime of biological systems. Note that a positive voltage of 2.5V is applied to the output terminal to aid the collection of the weight charge packet.

The surface potential profiles of the charge coupled synapse at different spike times are shown in Fig. 3.16. The profiles exhibit similar characteristics to the potential profiles shown in Fig. 3.12. Before the pre-synaptic spike is applied to the gate of MOS capacitor C2, the synapse is in equilibrium, and the potential barrier, as indicated in Fig. 3.16, hinders the charge transfer from the storage well to the output terminal. As soon as the spike from pre-synaptic neuron arrives at C2, the deeper

potential well is formed in the silicon of C2 and the potential barrier vanishes, allowing the weight charge to flow towards the output terminal. It is clearly shown in the figure that the potential of C1 increases to its maximum at 2μs, as a result of the weight charge transfer.

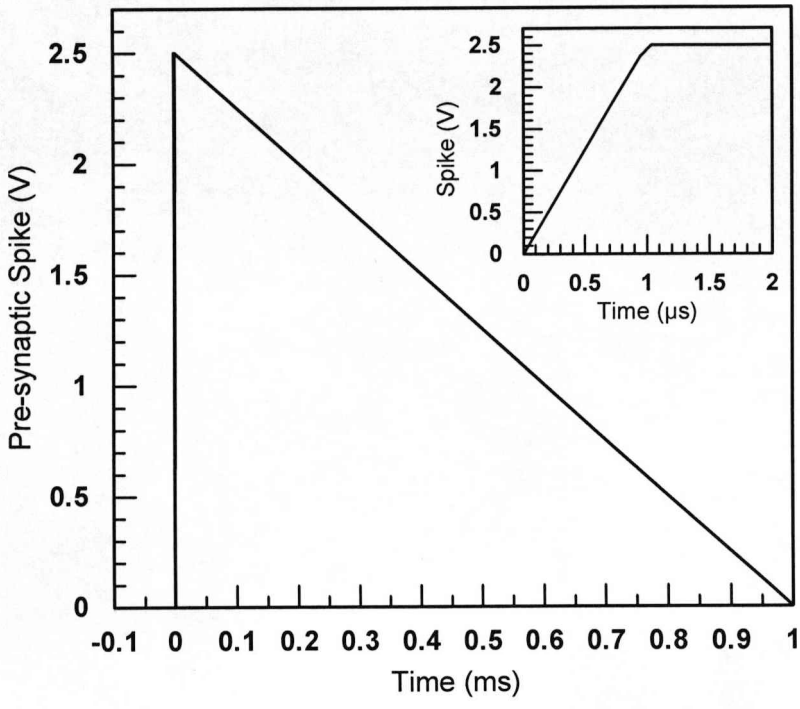


Fig. 3.15 Pre-synaptic spike V_i to the gate of the MOS capacitor C2. The spike has a rise time of 1μs, a duration of 1μs, as indicated in the inset, and a 1ms fall time.

Fig. 3.17 shows the weight charge concentration under the gate of C1 over spike time. The decrease of the charge concentration follows the increase of the pre-synaptic spike. As indicated in the inset, the weight charge packet with the concentration of $2\times10^{16}cm^{-3}$ is left in the storage well of C1 at 0.6μs, and the charge transfer efficiency achieves 98% at 0.8μs. After the charge transfer process, the MOS capacitor C1 starts to re-establish the weight charge packet Q_W, as shown in Fig. 3.17, by a slow thermal generation process due to the weight voltage V_{ji}. When the pre-synaptic spike reaches 0V at 1ms, the storage well has not been fully refilled, and the surface potential of C1 has not reached its equilibrium value, as shown in Fig. 3.16. This non-equilibrium process will be discussed further in Chapter 4.

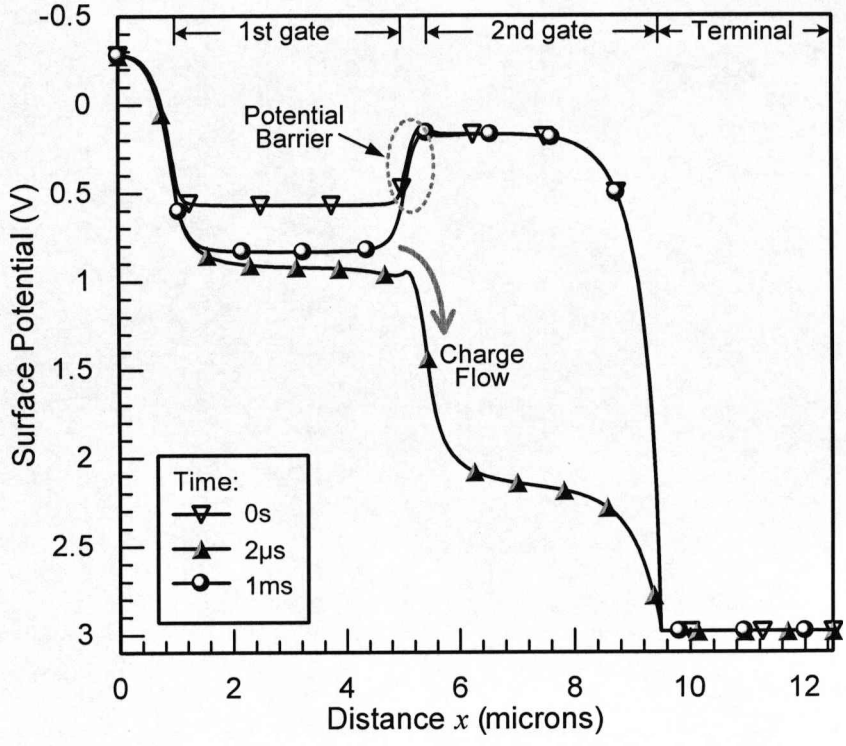


Fig. 3.16 Potential profiles along the surface of charge coupled synapse at 0s, 2μs, and 1ms.

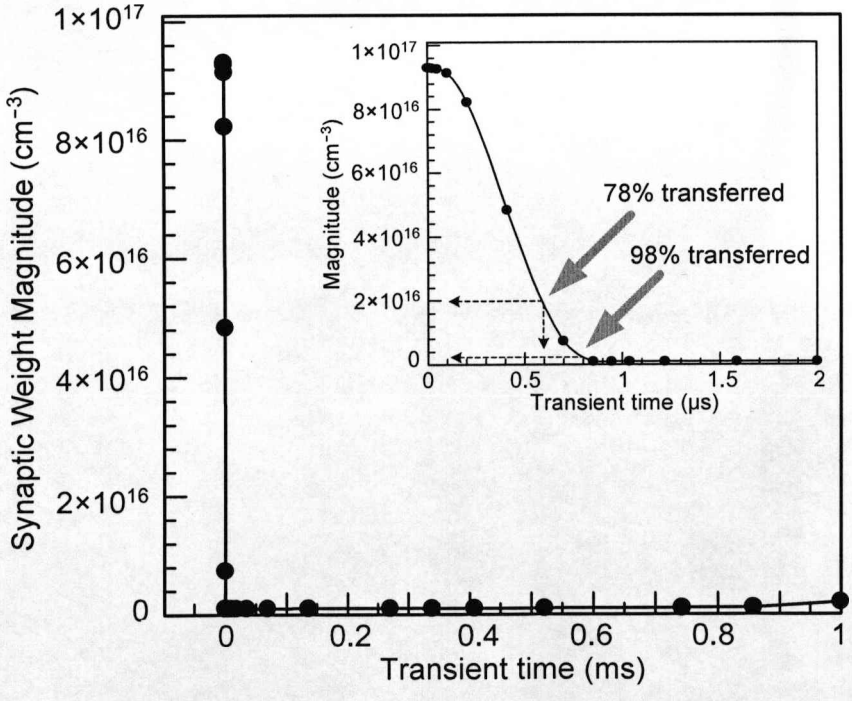


Fig. 3.17 Electron concentration under the gate of MOS capacitor C1 over pre-synaptic spike time. The inset shows the drop of the electron concentration in the first 2μs time.

Section 3.6.3 Spiking Output Current

(1) Pre-synaptic Spike with 1ns Rise Time

The overall time-dependency of the resulting current due to the weight charge transfer at the output terminal is shown in Fig. 3.18. When the electron concentration drops to $2.9\times10^{15}cm^{-3}$ where the thermal diffusion constant equals the effective drift constant, the charge transfer will be dominated by thermal diffusion. Therefore, it can be seen from the inset of Fig. 3.18 that for times less than 2ns, the charge decay is dominated by self-induced drift and the transfer efficiency achieves over 87%. The drift dominated time is shorter than the theoretical value since the fringing field greatly speeds up the transfer process even though the ramp stage is responsible for a small amount of charge decay. Further decay due to thermal diffusion is exponential with constant $\tau_{TD} = 6ns$ for the remaining charge.

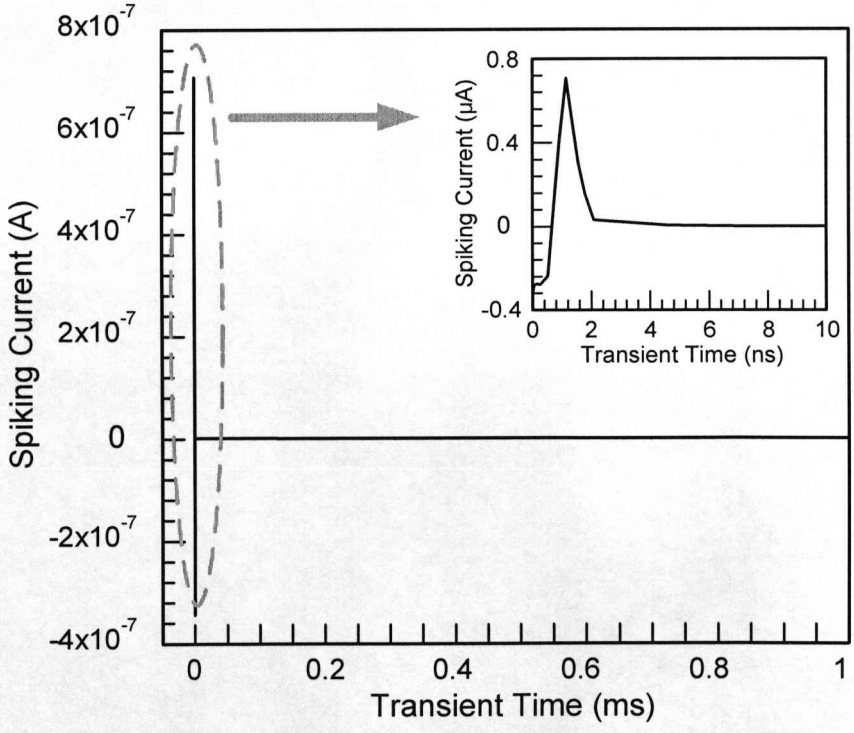


Fig. 3.18 Spiking current at the output terminal. The inset shows the details of the spike due to the arrival of the weight charge packet.

With the fast-rising spike on the gate of C2, a displacement current is caused by the *dV/dt* across the coupling capacitor between the gate of C2 and the output terminal. Therefore the current initially goes negative as indicated clearly in the inset of Fig. 3.18. With increasing transient voltage on C2, the potential gap disappears and the electrons flow towards the deep potential well causing the current rise. Over time, the remaining charge decays and the charge arriving at the output terminal diminishes. Therefore, after 1ns the spiking current starts to decline. The result shows that the spike characteristic of the proposed charge coupled synapse depends on the charge transfer mechanism.

To investigate the relationship between the spiking current and the weight voltage V_{ji} on the gate of C1, a set of simulations with the weight voltage V_{ji} varied from 1V to 2V has been run. Other parameters used in the simulations are the same as before. The spiking current characteristics shown in Fig. 3.19 indicate that the peak value is 1×10^{-6}A at 1ns for 2V weight voltage. For the weight voltage V_{ji} of 1.5V, the amplitude of the current is 9×10^{-7}A.

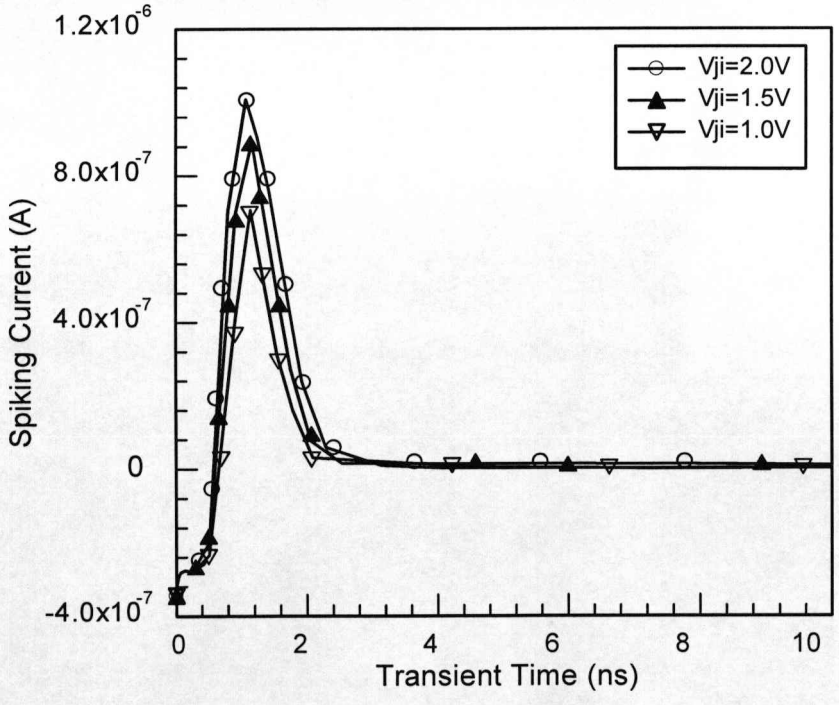


Fig. 3.19 Current spikes on output node for different weight voltage V_{ji} on the MOS capacitor C1. 100nm gate oxide thickness, 4μm electrode length, and 0.5μm electrode spacing; time step around the spikes is the order 1ps.

Fig. 3.20 shows the results of simulations with 25nm oxide thickness, 1μm electrode length and electrode spacing equal to 0.25μm. In Fig. 3.20 the weight voltage V_{ji} on C1 is varied from 0.4V to 1V in steps of 0.2V. Current spikes of different magnitude are observed at the output terminal where the integral of the spike curve is equal to the proportion of weight charge transferred. Note that the variation of the spike current amplitude with V_{ji} is, to a good approximation, linear in agreement with conventional MOS physics and this relationship is used to implement synaptic plasticity in the proposed charge coupled synapse. The slight initial undershoot arises from the displacement current due to charge coupling between the gate of MOS capacitor C2 and the output terminal. Fig. 3.21 shows the correlation between current spike amplitude and weight charge concentration which again supports plasticity in the proposed charge coupled synapse.

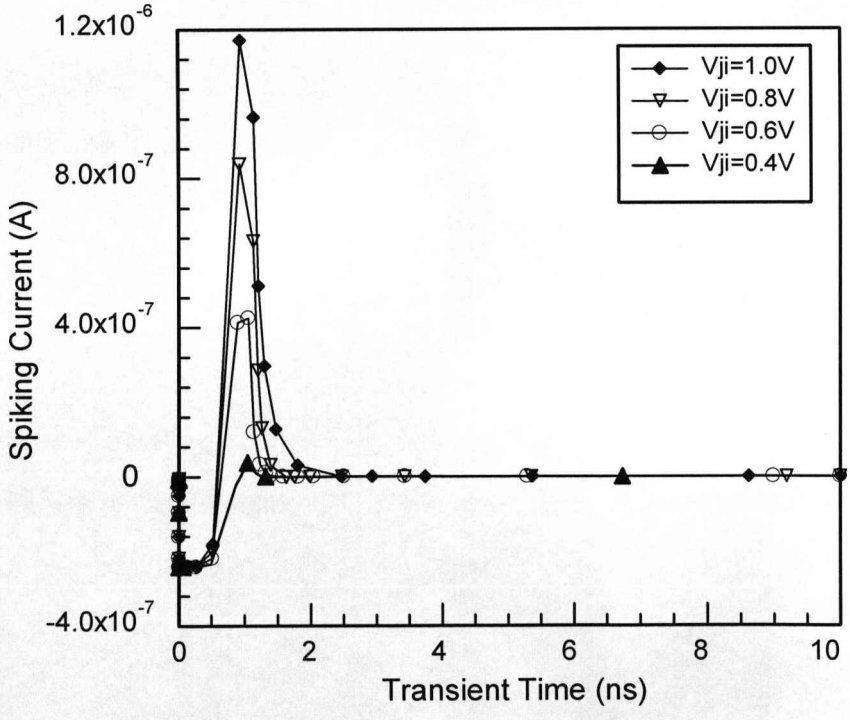


Fig. 3.20 Current spikes on output node for different weight voltage V_{ji} on the MOS capacitor C1. 25nm gate oxide thickness, 1μm electrode length, and 0.25μm electrode spacing; time step around the spikes is the order 1ps.

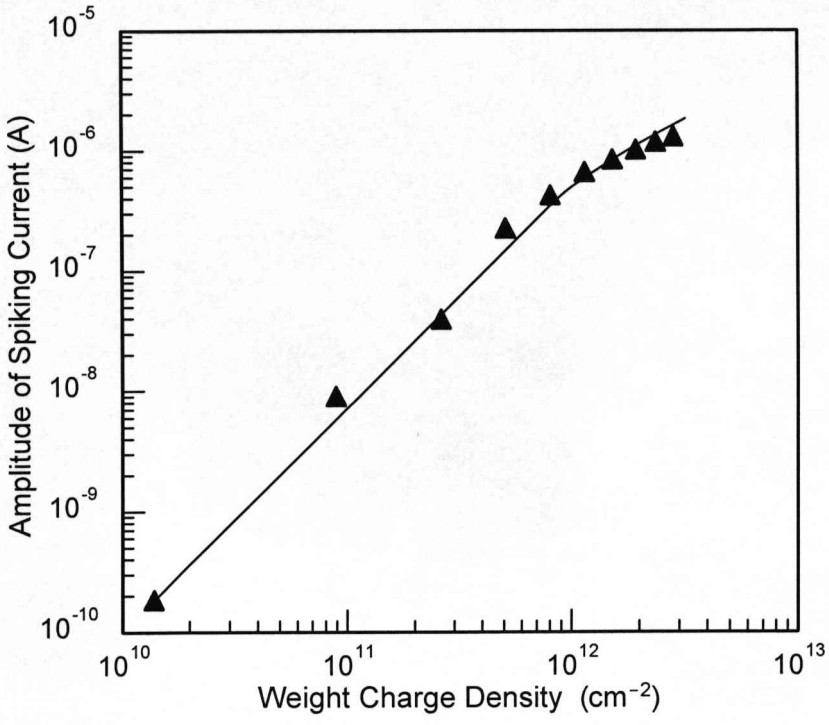


Fig. 3.21 Current spike amplitude versus weight charge density. 25nm gate oxide thickness, 1μm electrode length and 0.25μm electrode spacing.

(2) Pre-synaptic Spike with Slow Rise Time

The spiking current is generated at the output terminal of the charge coupled synapse due to weight charge transfer, in response to the pre-synaptic spike shown in Fig. 3.15. The current transient, similar to the characteristic of the biological spike, is shown in Fig. 3.22. The increase of the spiking current follows the increase of the pre-synaptic spike, and the amplitude of the current is about 2.5nA. Fig. 3.23 shows the spiking current with the amplitude of 0.15μA, in response to a pre-synaptic spike with 10ns rise time. It should be noted that the amplitude and duration of the simulated pre-synaptic spike could be varied to produce expected signals which are going to be fed into the subsequent neuron cells. In addition, the fall time of the spiking current is partly determined by the leakage current in the n^+-p junction formed by the N+ implant and p-substrate, which depends on the electron/hole Shockley-Read-Hall generation lifetime.

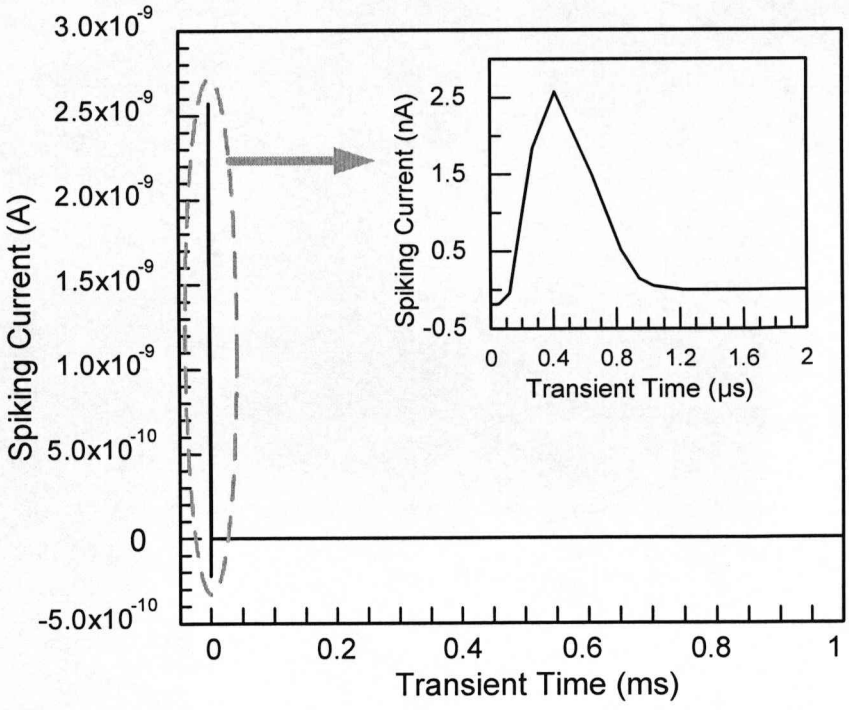


Fig. 3.22 Spiking current at the output terminal in response to the pre-synaptic spike with 1μs rise time and 1ms fall time. The inset shows the details of the current spike.

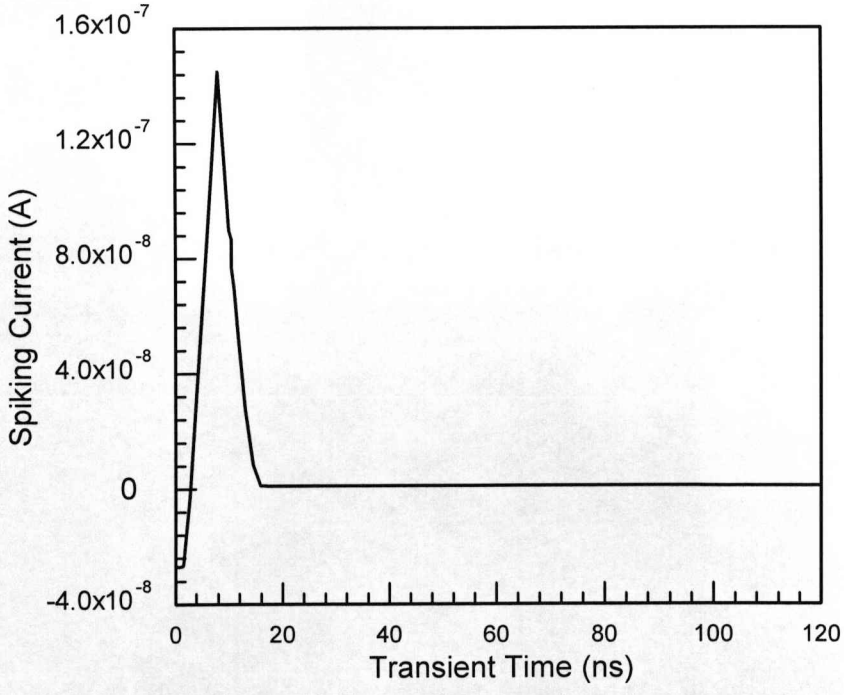


Fig. 3.23 Detailed spiking current at the output terminal in response to the pre-synaptic spike with 10ns rise time and 1ms fall time.

Section 3.6.4 Post-Synaptic Potential

The synapse structure with floating diffusion output stage, described in Section 3.5, is simulated to demonstrate the synaptic behavior as for a biologically plausible spiking neuron cell. The length of the output terminal is 1μm. The fixed positive voltage on the output terminal has been removed. The pre-synaptic spike, shown in Fig. 3.15, is applied to the gate of the MOS capacitor C2. This induces a positive potential on the floating diffusion node: a self-induced virtual bias. The similar synaptic weighting process is observed in the simulation. When the pre-synaptic spike arrives at the synapse, the potential barrier disappears and a deeper potential well is formed. The weight charge then flows to the MOS capacitor C2 and subsequently to the output terminal. After the charge transfer process, the charge coupled synapse starts to recover to equilibrium.

Fig. 3.24 shows the time dependence of post-synaptic potential (PSP) at the floating diffusion output region with the generation lifetime parameter set to 10ns. Due to the capacitive coupling between the second gate and the output terminal, the potential at the output increases sharply which strictly follows the increase of the pre-synaptic signal V_i, and this mimics the rise of the PSP in real neuron cells. As the lateral potential barrier between two electrodes disappears and the electrons flow onto the output terminal, the increase of the output potential is inhibited. The effect of the arrival of the weight charge packet is to reduce the overall terminal potential, causing a small spike as indicated in the inset of Fig. 3.24. The further relaxation of the potential, with much longer time constant, is caused by the falling of the pre-synaptic spike and the lifetime dependent leakage of the output/substrate diode. Then the potential becomes negative and resets to its equilibrium value slowly. Therefore the charge coupled synapse weights the pre-synaptic signal and delays it by an amount depending on the generation lifetime. Note that the decay time constant also depends on the carrier generation rate, that is strongly temperature dependent due to the strong temperature dependence of the intrinsic carrier density, which may not be a major problem given the intended very low operating power levels of the full system.

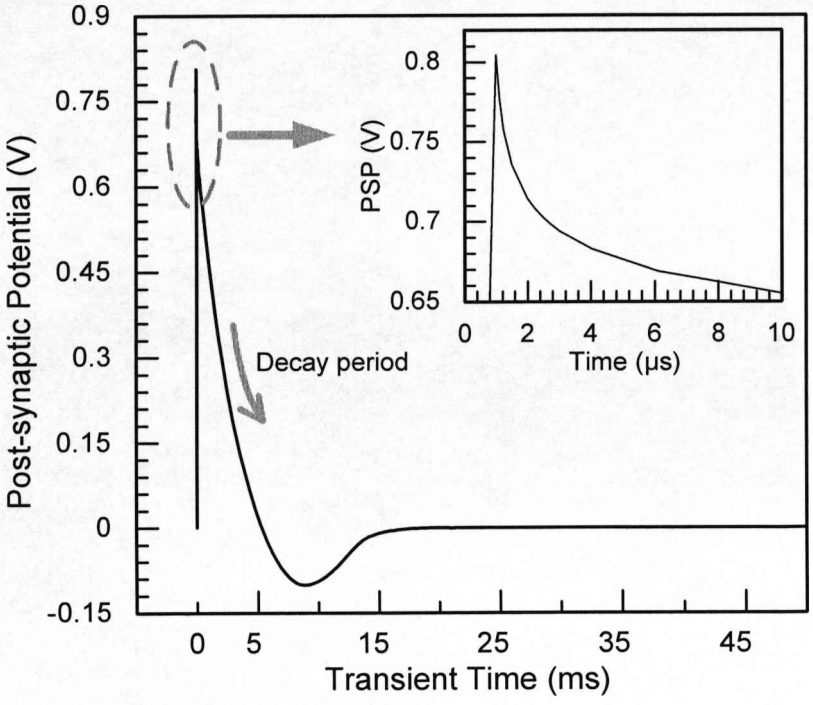


Fig. 3.24 Post-synaptic potential (PSP) at the floating diffusion output stage over time. Generation lifetime is 10^{-8}s.

A set of simulations has been carried out to investigate the lifetime dependent leakage effect of the output/substrate diode. The effect due to the falling of the pre-synaptic spike is not considered in the simulation. Fig. 3.25 shows the time dependence of PSP with different generation lifetimes. The relaxation of the PSP, with much longer time constant, is caused by the lifetime dependent leakage of the output/substrate diode on the order of milliseconds. As shown in Fig. 3.25, the latency of the FDN potential decreases with the decreasing lifetimes. This effect verifies the analytical model described in (3.35) and mimics the decay of the PSP in the realistic synaptic operation.

The fall times of the PSPs, which are proportional to the generation lifetime τ_g as described by (3.35), are shown in Fig. 3.26. For the 1ns lifetime, the latency is about 0.67ms; whereas the latency increases to 6.15ms for the lifetime of 10ns. This linear relation provides a possible engineering solution for the biologically plausible synapse in silicon.

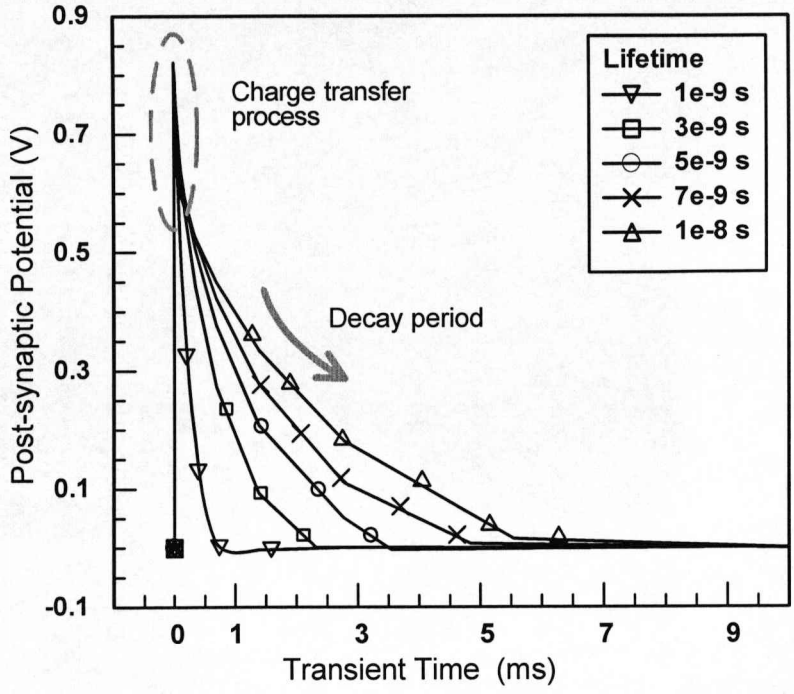


Fig. 3.25 The lifetime dependent FDN potential mimicking the biological PSPs due to the activation of the synapse.

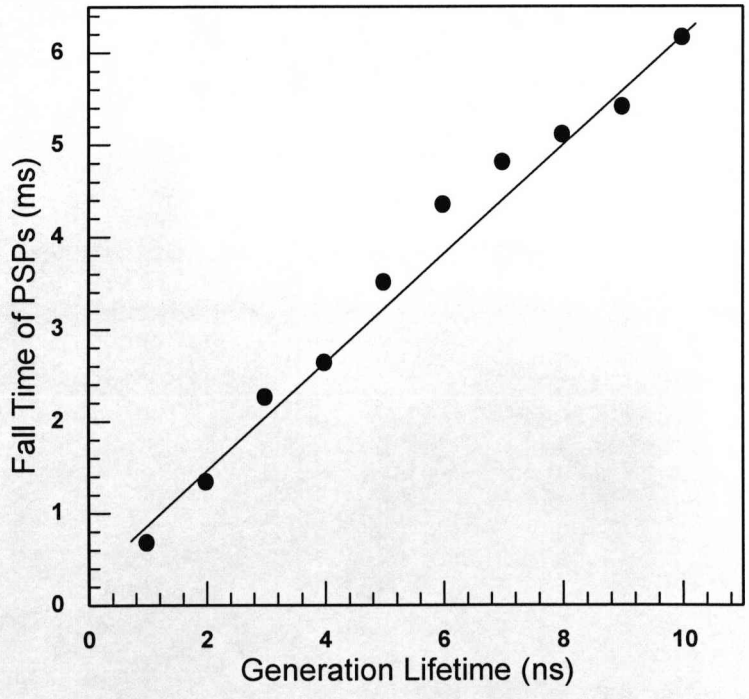


Fig. 3.26 The decaying times of the PSPs for different generation lifetimes providing a solution for engineering the biological plasticity.

The amplitude of the PSPs is directly proportional to the weight charge packet Q_W as shown in Fig. 3.27. Note that the charge density is the reflection of the weight voltage to the first MOS capacitor C1. In weak inversion mode, the small amount of charge has little effect on the output spike. With the same coupling capacitance C_{ov}, the amplitude of the PSPs examined at the floating diffusion terminal decreases linearly with the modulation as indicated in Fig. 3.27. Therefore, it is verified that the functionality of synapses in spiking neuron cells is effectively realized by the charge coupled synapse, and the nature of the signal transfer due to capacitive coupling leads inherently to low power operation.

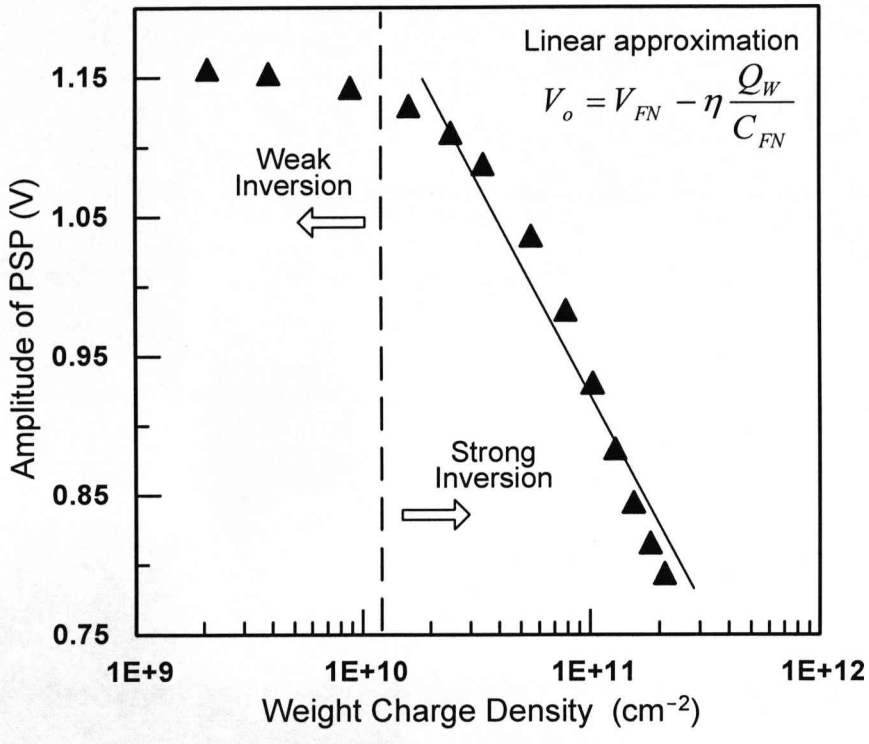


Fig. 3.27 Amplitude of PSPs versus weight charge density in the first MOS capacitor C1.

Section 3.7 Discussion and Conclusions

In this chapter, a compact charge coupled synapse exhibiting spiking behavior was developed as a core building block for spiking neural networks. The proposed silicon

synapse is based on the charge transfer device with associated localized memory capability integrated into a floating gate on the first phase. An analytical model for the lifetime dependent characteristic has been presented. Simulations were carried out to verify the plausibility of the synapse. By tuning the generation lifetime in the charge coupled synapse, the spiking behavior of the proposed synapse closely mimics the dynamics of realistic synapses in biological systems.

To illustrate the packing density afforded by the proposed charge coupled synapses, consider a conservative area estimate of 5μm×5μm ($=25\times10^{-12}m^2$) for the learning synapses. Using this estimate, it is easy to compute that the density of synapses/cm^2 is 4 million. This is only a rough estimate as inter-neuron interconnects and associated circuitry would have to be accounted for in the calculation. Nevertheless, if we assume that the total chip area is $1cm^2$ and make a conservative guess that only 10% of the total surface area is occupied by synapses, then it is easy to compute that 400 thousand synapses for a single layer planar process is possible, as compared to a few thousand which is the current estimate reported in the literature [9].

To estimate the worst case dynamic power consumption at network level, consider a network containing a million synapses where each synapse is operated at a frequency of 100Hz and assume zero phase differential across the network: 100Hz is typical of real neurons. Also assume a maximum weight voltage V_{ji} on each synapse of 2.5V which results, upon charge transfer, in a current spike with average amplitude of 1μA over a duration of 4ns. Hence, the total dynamic power consumed per second across the entire network of synapses is 1μW. Clearly this is negligible and will be significantly less than the power consumption due to the interconnect bus architecture and associated decoders etc.

The charge coupled synapse is demonstrated to be able to capture the intrinsic dynamics of the real synapse and so can mimic the synaptic plasticity in a spiking neuron cell. The charge coupled synapse possesses the potential for scalability, associated with engineering the post-synaptic neuron circuits, to produce biologically plausible spiking neural networks in hardware, advancing the exploitation of brain-like computational systems.

References

[1] Y. Taur and T. H. Ning, *Fundamentals of modern VLSI devices*, Cambridge University Press, 1998.

[2] J. E. Carnes, W. F. Kosonocky, and E. G. Ramberg, "Drift-aiding fringing fields in Charge-Coupled Devices," *IEEE Journal of Solid-State Circuits*, vol. SC-6, no. 5, pp. 322–326, Oct. 1971.

[3] J. E. Carnes, W. F. Kosonocky, and E. G. Ramberg, "Free charge transfer in Charge-Coupled Devices," *IEEE Transactions on Electron Devices*, vol. ED-19, no. 6, pp. 798–808, June 1972.

[4] C. Kim, "The physics of Charge-Coupled Devices," in *Charge-coupled Devices and Systems*, M. J. Howes and D. V. Morgan, Eds. Chichester, UK: John Wiley & Sons, 1979, 1–80.

[5] W. E. Engeler, J. J. Tiemann, and R. D. Baertsch, "Surface charge transport in silicon," *Applied Physics Letter*, vol. 17, no. 11, pp. 469–472, December 1970.

[6] C. K. Kim, "Carrier transport in charge-coupled devices," *Proceeding of ISSCC*, Philadelphia, Pa, 1971.

[7] W. E. Engeler, J. J. Tiemann, and R. D. Baertsch, "A memory system based on surface-charge transport," *Proceeding of ISSCC*, Philadelphia, Pa, 1971.

[8] W. F. Kosonocky and J. E. Carnes, "Charge-coupled digital circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-6, no. 5, pp. 314–322, Oct. 1971.

[9] E. Chicca, D. Badoni, V. Dante, et al., "A VLSI recurrent network of integrate-and-fire neurons connected by plastic synapses with long-term memory," *IEEE Trans. Neural Networks*, vol. 14, no. 5, pp. 1297–1307, Sep. 2003.

[10] D. K. Schroder, "Carrier lifetimes in silicon," *IEEE Trans. Electron Devices*, vol. ED-44, no. 1, pp. 160–170, 1997.

[11] T. Shibata and T. Ohmi, "A functional MOS transistor featuring gate level weighted sum and threshold operations," *IEEE Transactions on Electron Devices*, vol. 39, no. 6, pp. 1444–1455, June 1992.

CHAPTER 4 PROGRAMMABLE CHARGE COUPLED SYNAPSE

Section 4.1 Introduction

Biological spiking neurons communicate with each other primarily through fast chemical synapses. By converting trains of action potentials into varying amplitudes of post-synaptic responses, synapses perform a type of temporal filtering [1]. Consider a fragment of spiking neural networks (SNNs) [2] consisting of two point neurons with a connecting synapse, as shown in Fig. 4.1. Synaptic transmission has been shown to be a dynamic process. A fundamental property of synapses is their ability to modify the efficacy of synaptic communication through various forms of synaptic plasticity. Virtually all types of synapses, which exhibit a range of use-dependent behavior, are regulated by a variety of short-lived (tens of milliseconds) and long-lasting (several hours or days) processes. Recent experiments have revealed that at the synaptic junction post-synaptic responses are not simply a function of the pre-synaptic firing rate multiplied by a synaptic weight but rather, reflect the short-term history of input spike trains. Since the paired firing of pre- and post-synaptic neurons tends to redistribute total synaptic efficacy, a synapse responds much more strongly to the first pre-synaptic event than to subsequent events in a spike train [3]. Short-term plasticity is believed to be a crucial factor in the processing of information in spiking neural networks [4]. There are two types of short-tem plasticity actions which are referred to as depressing or facilitating and phenomenological models have been proposed to describe both of these transmission states [5]. With repeated use, synaptic enhancement can occur and facilitation dominates while at other synaptic sites the result is a decrease in synaptic strength and depression prevails. In some cases multiple processes are present and the result can be a combination of facilitation and depression [6] in which synaptic strength is highly dependent on the details of the timing of synaptic activation. These biological characteristics have attracted the attention of neuromorphic engineers who focus mainly on the dynamic implications of the neurons. Recently work has begun on the implementation of dynamic synapses in hardware where an analogue approach was presented [7]. It demonstrated that a

neural network can achieve pattern recognition and moreover its performance can be improved by using depressing synapses. In other work [8] part of a multi-chip system implemented a Winner Take All (WTA) network which used dynamic synapses and adaptive neurons. An Address Event Representation (AER) communication system allowed interfacing to silicon spiking retinas and to software implementations of associated memories.

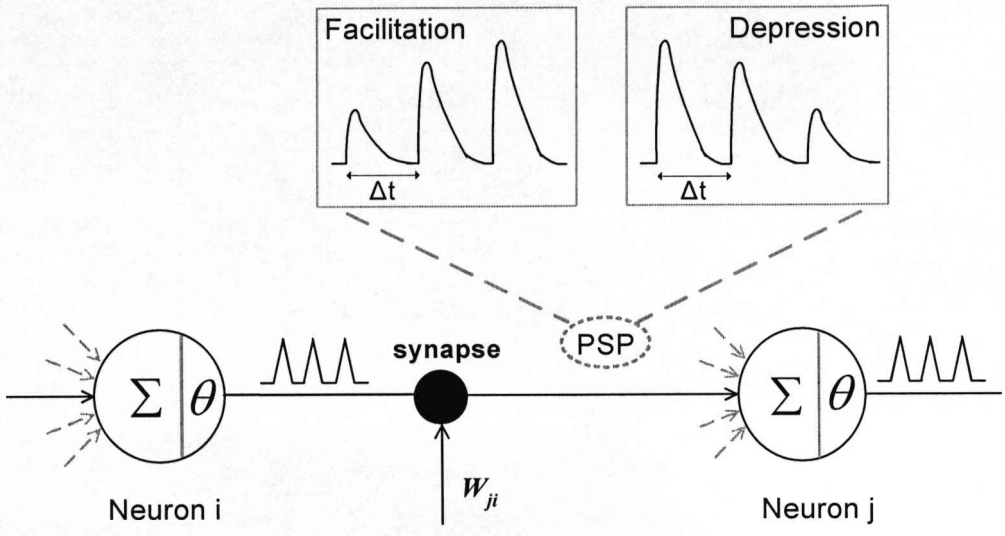


Fig. 4.1 Illustration of a fragment of SNNs with synaptic junction. θ represents the threshold. The insets show the facilitation and depression behavior at many synapses with stimuli separated by time Δ*t*.

In Chapter 3, a charge coupled synapse was developed, which forms the basis of the work presented in this chapter, to mimic the spiking behavior of the biological synapse for the hardware implementation of spiking neural networks. The weight charge packet Q_W stored in the well of a MOS capacitor is released by the pre-synaptic spike dropped onto the second MOS capacitor. Each time the storage well is emptied of charge, a weighted current/voltage spike is produced at the output terminal where the weighting is set by the equilibrium charge density in the well: the linear relationship between the weight charge density and the associated gate voltage is used to implement plasticity. A disadvantage of the approach was the dependency of the dynamic response on the generation lifetime which is process dependent. The device investigated in this chapter is designed to avoid this dependency by allowing control of the relaxation by an injecting junction that can be tuned appropriately.

In this chapter, the recovery of Q_W by thermal generation of electron-hole pairs is mathematically modeled, and the response of a charge coupled synapse to successive pre-synaptic spikes is investigated. However this slow process fails to match the time scale of the inter-spike interval (ISI) of biological system which can be as low as a few milliseconds. Therefore the device concepts for the programmable dynamic charge coupled synapses, where the relaxation process can be greatly accelerated, are presented. The first programmable dynamic synapse described in Section 4.4 is based on the charge coupled synapse to which an additional source of minority carriers is attached. The programmable functionality of the synapse is implemented by the weight restoration through charge injection from an N+ implant pulsed by a small negative voltage. Correspondence is made between the semiconductor relaxation processes and biologically relevant responses. An alternative type of programmable dynamic charge coupled synapse presented in Section 4.5, consists of an injector MOS transistor in proximity to two MOS capacitors. The weight is stored as a package of charge in the well of a MOS capacitor and the level of charge in the well is modulated by the associated gate voltage. Successive spikes or a spike train from the pre-synaptic neuron is applied to the gate of the nearby capacitor and repeatedly empties the storage well of charge. After each spike event the well is filled by charge arising from an injector MOS transistor, which operates in the subthreshold current regime. The magnitude of this current, which is controlled by the gate voltage, sets the minimum ISI associated with the pre-synaptic train. The device developed within this section is designed in the framework of a standard mixed signal CMOS process from Austria MicroSystems. Simulation results are presented to justify the assertion that the proposed silicon synapses captures the dynamics of a biological synapse by using innate features of the semiconductor physics.

Section 4.2 presents the transient analysis of the thermal generation process. Section 4.3 describes the non-equilibrium response of a charge coupled synapse to a pre-synaptic spike train. The charge coupled synapse with a charge injector is proposed and investigated in Section 4.4 together with the simulation study. Section 4.5 presents the operating principles and dynamic behavior of the three-phase charge coupled synapse while Section 4.6 concludes the chapter.

Section 4.2 Transient Operation of Weight MOS Capacitor

As described in the previous chapter, the charge coupled synapse shown in Fig. 4.2 comprises two MOS capacitors in proximity, one of which has charge storage capability. The synaptic weight is therefore represented by the stored charge packet Q_W. Immediately after the charge Q_W is released by the pre-synaptic spike V_i and transferred, the charge coupled synapse is in a state of non-equilibrium with deep depletion conditions under the gates of two MOS capacitors. Under this condition the overall relaxation to equilibrium re-establishes the inversion layer charge Q_W, through the generation of electron/hole pairs in the inversion layer beneath the gate of C1. Therefore the complex synaptic plasticity is realized by the adaptation of the charge volume depending critically on the timing of the spikes from both pre-synaptic and post-synaptic neurons. Since the timing of each relative spike can be as little as 10ms, which is considerably faster than the time for a MOS capacitor to reach equilibrium, time-dependent thermal generation emerges as a key mechanism underlying various forms of synaptic plasticity.

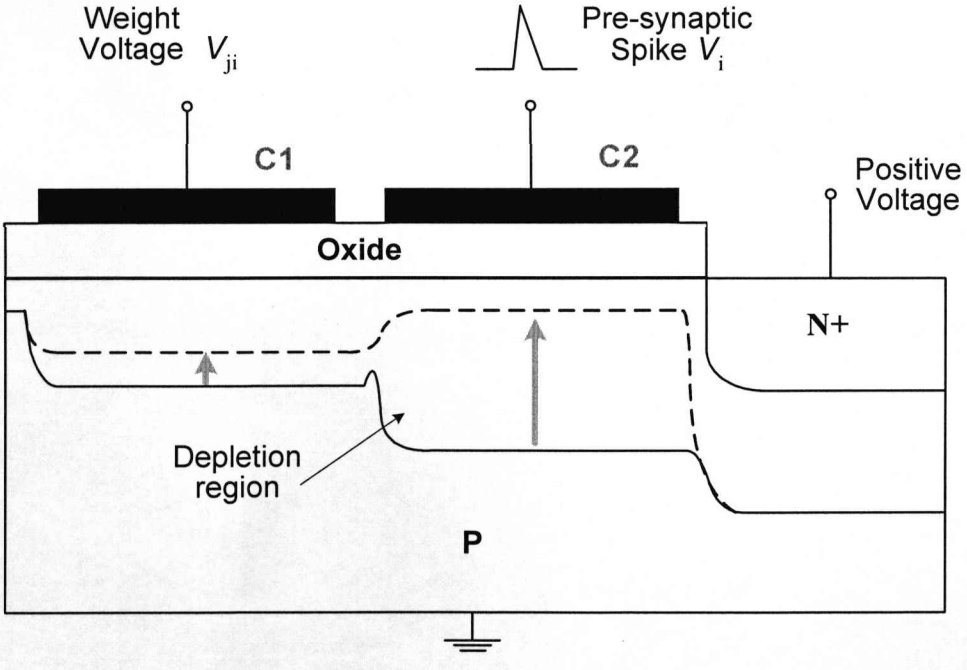


Fig. 4.2 Schematic view of a charge coupled synapse under non-equilibrium state with deep depletion conditions. The depletion width goes back to the equilibrium value (dashed) after the charge transfer process.

It is now necessary to develop a physical model for the time-dependent process of weight generation in the first MOS capacitor C1. The energy band diagram of an ideal MOS capacitor after the application of a step voltage V_{ji} is shown in Fig. 4.3.

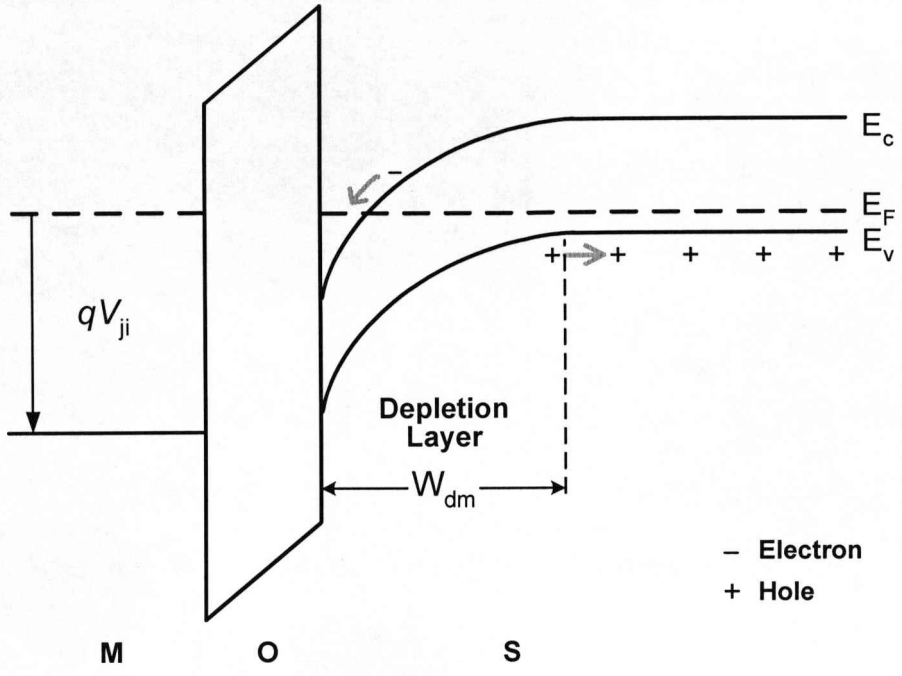


Fig. 4.3 Energy band diagram under deep depletion condition of the MOS capacitor C1. The depletion layer relaxes due to thermal generation.

If the MOS capacitor C1 is initially in the flatband condition, charge is neutral throughout the device. Immediately after the application of the step voltage V_{ji}, the surface potential is large and depends on the dielectric properties of the oxide and semiconductor. The majority carriers deplete very rapidly, within a dielectric relaxation time, establishing an extended depletion region. Thus the device is in the deep depletion state, and the acceptor charge within the depletion region compensates the gate charge. At this time the surface potential corresponds to the non-equilibrium depletion width. As time progresses, an inversion charge layer starts to build up causing the depletion region to retract. This relaxation of the semiconductor occurs by the generation of electron-hole pairs within the depletion region with each generated electron adding to the inversion layer. As time passes, the electrons at the surface

build up and the surface potential decreases further to its final value when the device is approaching equilibrium.

A mathematical model is now developed for the rate of change of inversion charge density with time in a MOS capacitor under non-equilibrium conditions. As described in Section 2.2.1 (see Eqn. (2.11)), the gate voltage V_{ji} equals the sum of the flatband voltage V_{FB}, the voltage across the oxide V_{ox} and the potential across the semiconductor ϕ_{s1}:

$$V_{ji} = V_{FB} + \phi_{s1} + V_{ox} = V_{FB} + \phi_{s1} - \frac{Q_{si}}{C_{ox}} \tag{4.1}$$

where C_{ox} is the oxide capacitance; Q_{si} is the total charge in semiconductor including both the depletion and inversion components. Fixed oxide and interface charge are considered to be incorporated into the flatband voltage and taken to have negligible influence on the transient physics of the MOS capacitor. Therefore, the gate voltage equation becomes:

$$V_{ji} = -\frac{Q_d + Q_{inv}}{C_{ox}} + \phi_{s1} = \frac{t_{ox}}{\varepsilon_{ox}\varepsilon_0}\left(qN_aW_{d1} + qn_{inv}\right) + \frac{qN_aW_{d1}^2}{2\varepsilon_{si}\varepsilon_0} \tag{4.2}$$

where t_{ox} is the oxide thickness; N_a is the substrate doping density; W_{d1} is the depletion width of C1; n_{inv} is the electron density in the inversion region; $\varepsilon_{ox}\varepsilon_0$ and $\varepsilon_{si}\varepsilon_0$ are the permittivities of SiO_2 and Si respectively. Differentiation of (4.2) gives:

$$\frac{dV_{ji}}{dt} = \frac{qN_at_{ox}}{\varepsilon_{ox}\varepsilon_0}\frac{dW_{d1}}{dt} + \frac{qt_{ox}}{\varepsilon_{ox}\varepsilon_0}\frac{dn_{inv}}{dt} + \frac{qN_aW_{d1}}{\varepsilon_{si}\varepsilon_0}\frac{dW_{d1}}{dt} \tag{4.3}$$

Since the gate voltage V_{ji} is constant during the transient, (4.3) becomes:

$$N_a\left(1 + \frac{\varepsilon_{ox}}{\varepsilon_{si}}\frac{W_{d1}}{t_{ox}}\right)\frac{dW_{d1}}{dt} + \frac{dn_{inv}}{dt} = 0 \tag{4.4}$$

The increase of electrons in the inversion layer due to the generation of electron-hole pairs can be reasonably well represented as:

$$\frac{dn_{inv}}{dt} = \frac{n_i}{2\tau_g}\left(W_{d1} - W_f\right) \tag{4.5}$$

where W_f is the final (equilibrium) depletion width; n_i is the intrinsic concentration; τ_g is the generation lifetime in the depletion region [9]. Rearranging (4.4) by substituting (4.5) and the initial condition $W_{d1}(0)=W_{dm}$ gives:

$$\left(1+\frac{\varepsilon_{ox}}{\varepsilon_{si}}\frac{W_f}{t_{ox}}\right)\ln\left(\frac{W_{d1}-W_f}{W_{dm}-W_f}\right)+\frac{\varepsilon_{ox}}{\varepsilon_{si}}\frac{(W_{d1}-W_{dm})}{t_{ox}}+\frac{tn_i}{2\tau_g N_a}=0 \tag{4.6}$$

Substitute (2.2) and (2.5) into (2.7) and assume zero flatband voltage, the initial depletion width W_{dm} can be found by solving the quadratic equation:

$$V_{ji}=qN_aW_{dm}\frac{t_{ox}}{\varepsilon_{ox}\varepsilon_0}+\frac{qN_aW_{dm}^2}{2\varepsilon_{si}\varepsilon_0} \tag{4.7}$$

and the final depletion width W_f is found for the case of the surface potential equal to twice the bulk Fermi potential which yields:

$$W_f=\sqrt{\frac{4\varepsilon_{si}\varepsilon_0 V_t}{qN_a}\ln\left(\frac{N_a}{n_i}\right)} \tag{4.8}$$

where V_t is the thermal voltage.

For ease of computation, an empirical expression for the depletion width as a function of time, $W_{d1}(t)$, can be found by solving (4.6) and fitting a polynomial equation to the solution. Note that the physical parameters of the device are taken as: $N_a=10^{15}cm^{-3}$; t_{ox}=20nm; ε_{si}=11.9; ε_{ox}=3.9. The applied voltage V_{ji} is taken as 1V. The initial value for the depletion width is found to be 1.09μm from solving (4.7) and the surface potential is 0.9V. For the case of 10^{-7}s lifetime, the depletion width is obtained:

$$W_{d1}(t)=1.09\times10^{-4}-8.14\times10^{-5}t+1.27\times10^{-4}t^2-1\times10^{-4}t^3$$
$$+3.95\times10^{-5}t^4-6.08\times10^{-6}t^5 \tag{4.9}$$

The expression of depletion width prior to the equilibrium state of MOS capacitor C1 with different lifetimes is plotted in Fig. 4.4. The fast relaxation of bulk substrate due to the thermal generation of electron-hole pairs requires short lifetime. As shown clearly, it takes 0.99s for the MOS capacitor with 50ns lifetime to reach equilibrium; while it takes 1.99s and 9.96s in the case of 100ns and 500ns lifetimes respectively.

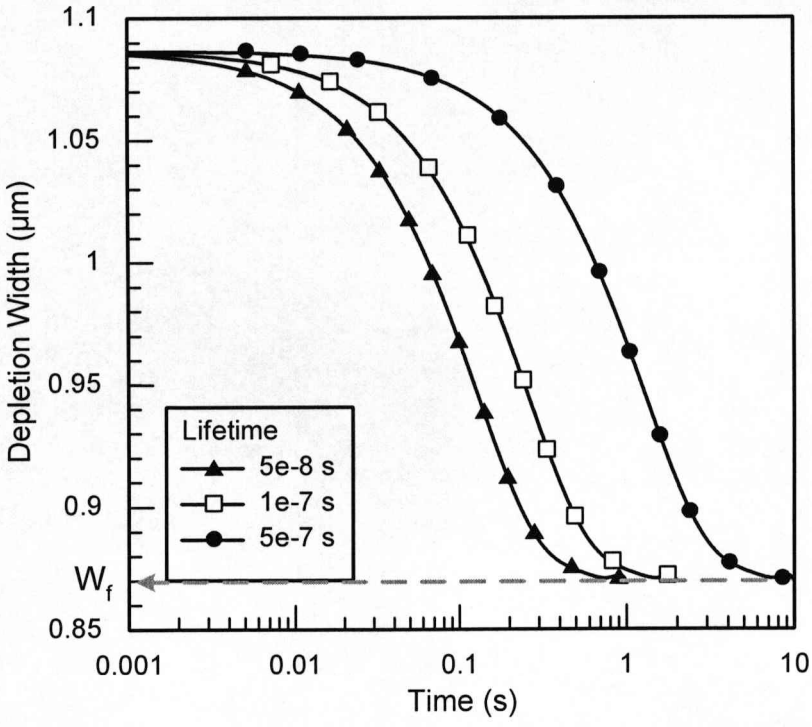


Fig. 4.4 Transient response of the depletion width in a pulsed MOS capacitor from flatband to equilibrium for different generation lifetimes. t_{ox}=20nm; N_a=10^{15}cm^{-3}; V_{ji}=1V.

Therefore, the rate of change of electron density in inversion layer can be obtained by substituting (4.9) into (4.5) and carrying out the integration:

$$n_{inv} = \frac{n_i}{2\tau_g}\int_0^t \left(W_{d1}(t) - W_f\right)dt \tag{4.10}$$

The response of the surface charge density n_{inv} prior to the equilibrium state of MOS capacitor is shown in Fig. 4.5. It can be seen that the minority carrier concentration rises almost linearly with time as if being supplied by a current source. At the time of 10ms, which is the timing of spike for realistic neurons, the inversion charge density n_{inv} is 3.1×10^{9}cm^{-2} in the case of 500ns lifetime. This approximates to a volume electron concentration of 2×10^{15}cm^{-3} in the storage well which is greater than the substrate doping density of 10^{15}cm^{-3}. For the lifetime of 10ns, the inversion charge density is 1.3×10^{11}cm^{-2} which approximates to a volume charge concentration of

$7.8\times10^{16}cm^{-3}$. The final inversion charge density is about $3.7\times10^{11}cm^{-2}$ as the MOS capacitor approaches equilibrium.

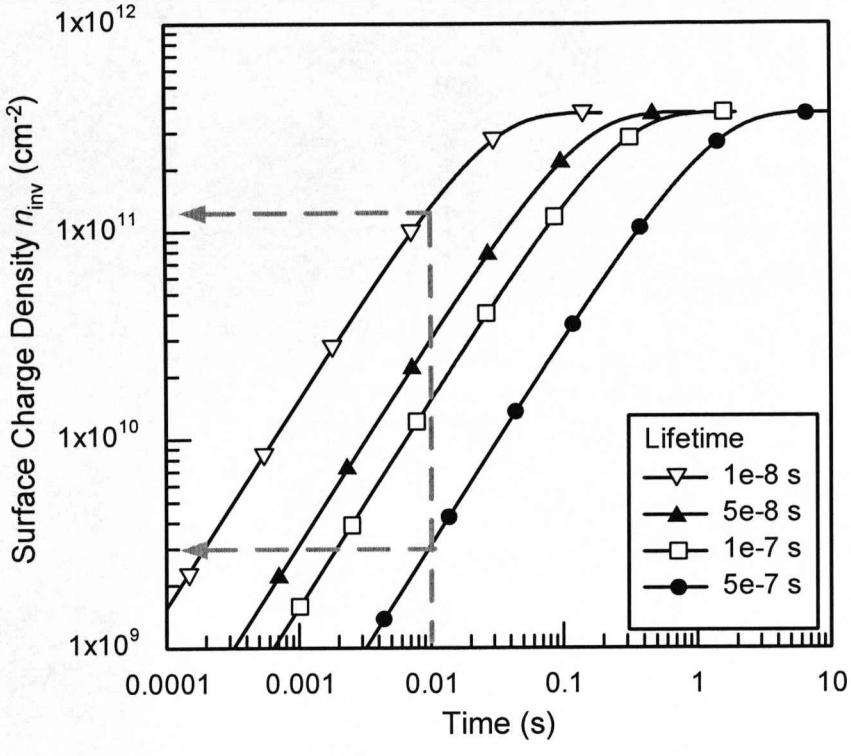


Fig. 4.5 Transient response of the surface charge density in a pulsed MOS capacitor from flatband to equilibrium. t_{ox}=20nm; $N_a=10^{15}cm^{-3}$; V_{ji}=1V. At 10ms, $n_{inv}=1.3\times10^{11}cm^{-2}$ for 10ns lifetime, and $n_{inv}=3.1\times10^{9}cm^{-2}$ for 500ns lifetime.

Fig. 4.6 clearly shows the relationship between the recovery time of the MOS capacitor C1 and the lifetime. This time-dependent relaxation mechanism described by (4.6) enables the implementation of dynamic synapses at device level in the biological time regime. The timing of the applied gate voltage, which depends on the pre-synaptic and post-synaptic spikes through a control circuit, can determine whether a synapse is potentiated or depressed. In addition, for the same time lag between pre-synaptic and post-synaptic spikes, the strength of potentiation and depression can be tailored to that expected for realistic neurons by setting the appropriate generation lifetime.

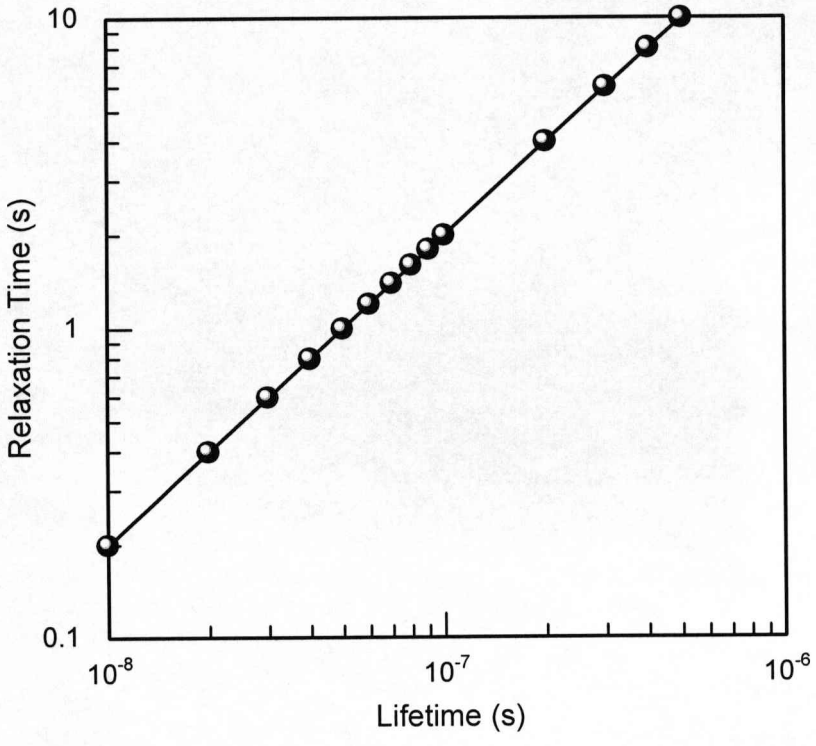


Fig. 4.6 Relaxation time of the MOS capacitor versus lifetime. t_{ox}=20 nm; N_a=10^{15} cm^{-3}; V_{ji}=1V.

Section 4.3 Transient Response to Successive Pre-synaptic Spikes

Fig. 4.7 shows the time-dependence of the weight charge density in the first MOS capacitor C1 for a series of pre-synaptic spikes with 1s ISI applied to the second gate of the charge coupled synapse. A fixed weight voltage (V_{ji}=1V) is applied to C1 at 0.5s such that the system is in equilibrium for times <0.5s. As soon as the first pre-synaptic signal arrives at the synapse at 1.5s, most of the charge Q_W is transferred rapidly (order of ns) to the neuron, leaving behind an empty storage well within the first capacitor. Note that at a time of 1.5s, the MOS capacitor C1 has not reached the equilibrium state and the inversion layer has not been completely formed, but the electron concentration is significantly greater than the doping density N_a. Immediately after the weight charge Q_W has transferred, the depletion layer width of C1 reverts back to the initial transient value W_{dm}, shown in Fig. 4.3, to balance the gate charge.

The MOS capacitor C1 is now driven into the deep depletion state again. The weight charge packet Q_W will be re-established by the generation of electron-hole pairs until it is released by the next pre-synaptic signal. As demonstrated in Fig. 4.7, the same amount of weight charge is generated each time for the same ISI. Considering an output terminal of the synapse shown in Fig. 4.2 of area $4\mu m^2$ and 2ns transfer time for the majority of the charge Q_W shown above, a spiking current with amplitude of 1.2μA will be induced and transmitted to the neuron cell circuit described in the next chapter.

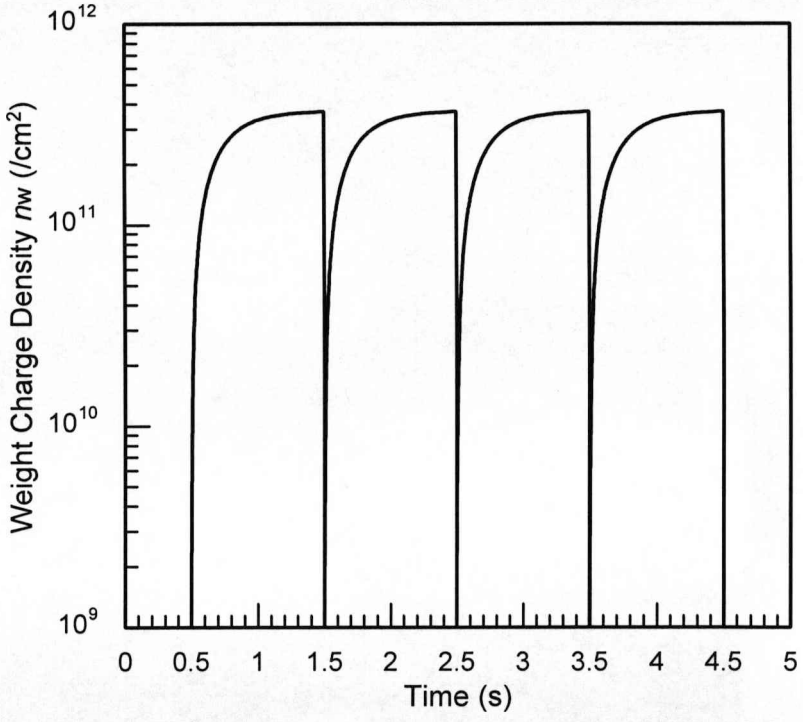


Fig. 4.7 Transient response of the weight charge density n_W to a series of pre-synaptic signals at 1.5s, 2.5s, 3.5s, and 4.5s. The MOS capacitor C1 is pulsed into non-equilibrium state. Parameters are: t_{ox}=20 nm; N_a=10^{15} cm^{-3}; V_{ji}=1V; τ_g=10^{-7} s.

For a pre-synaptic spike frequency of 100Hz (10ms ISI), the earlier analysis shows that the MOS capacitor C1 will not have been recovered completely. The associated weight charge density n_W, shown in Fig. 4.8, is therefore much less than that with 1Hz frequency shown in Fig. 4.7. Other parameters used in the calculation are the same. The plot clearly demonstrates the depression behavior of biological synapses. Each

time the pre-synaptic signal arrives at the MOS capacitor C2, the weight charge of the density $1.6\times10^{10}cm^{-2}$ is released and the resulting spiking current to the neuron circuit is of the amplitude of $5\times10^{-8}A$.

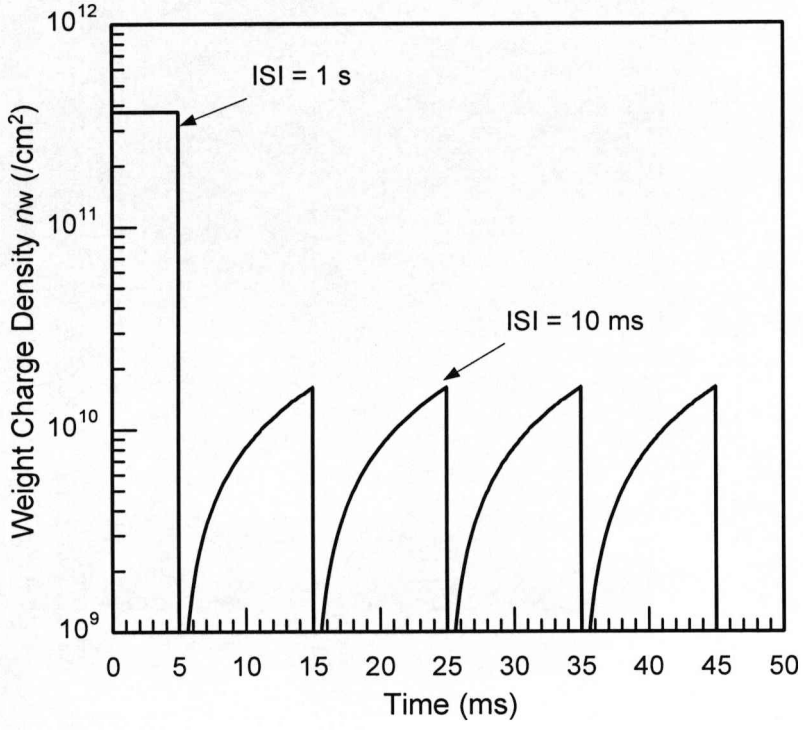


Fig. 4.8 Transient response of the weight charge density n_W to a series of pre-synaptic signals with 10ms ISI, soon after the release of large amount of charge due to 1s ISI. Parameters are: t_{ox}=20 nm; N_a=10^{15} cm^{-3}; V_{ji}=1V; τ_g=10^{-7} s.

Section 4.4 Programmable Synapse with Charge Injector

As analyzed in the previous sections, the thermal generation process in the MOS capacitor C1 is time-dependent and relies on the minority carrier generation lifetime, τ_g with a total relaxation time, t_{relax} given as $t_{relax} \sim \tau_g(N_a/n_i)$. For good quality silicon τ_g is of the order 10-100μs and therefore it will take many seconds to re-establish Q_W. If single spike encoding is used then recovery times of this magnitude may be acceptable. However, biological neurons emit and receive spike trains where the ISI can be as low as a few milliseconds, so a means to accelerate the relaxation process is

required. The reliance on the process-dependent carrier lifetime is also undesirable. The requirement can be addressed by providing an additional source of minority carriers into the substrate of the synapse. As shown in Fig. 4.9 the charge coupled synapse is modified to include an additional N+ implant, referred to as a 'charge injector'. Note that the lateral region beyond the N+ implant is not gated and can be controlled by masking to multiples of minimum feature size.

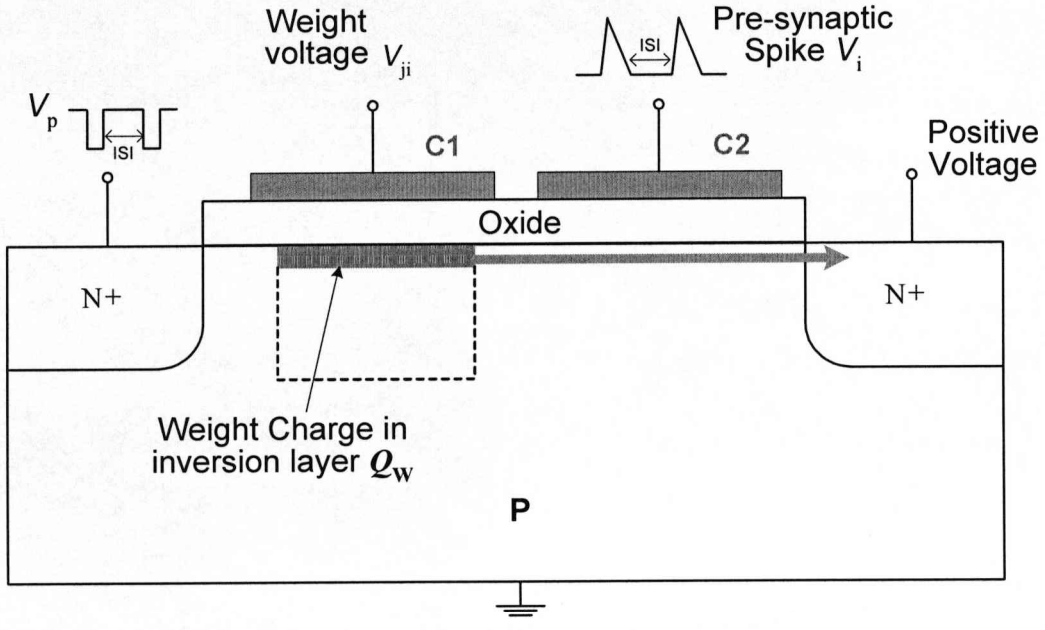


Fig. 4.9 Schematic view of the charge coupled synapse with a charge injector. The application of V_p is delayed by a time T relative to the pre-synaptic spike applied to C2.

The newly developed charge coupled synapse then, is based on a two-phase charge coupled synapse, essentially comprising of two MOS capacitors in close proximity. As described in Chapter 3, the synaptic weight is represented by the charge Q_W stored in the first MOS capacitor C1, which depends on the weight voltage V_{ji}. The signal emitted by a pre-synaptic neuron goes to the gate of the second MOS capacitor C2, causing the weight charge Q_W to be released which drives C1 into the deep depletion state. To remove the dependence on thermal generation in re-establishing Q_W, a small negative voltage pulse V_p is applied to the N+ implant, momentarily forward biasing the n^+-p junction. Under this condition, the p-substrate is flooded with minority

carriers (electrons) which are collected by the depletion layer edge of C1, and Q_W is quickly re-established. To ensure that the minority carriers in the substrate do not corrupt the spike current magnitude, and hence the weighting, the application of V_p is delayed by a time T relative to the pre-synaptic spike applied to C2, which sets the ISI of the synapse and consequently the operating frequency.

The change in electron concentration at the p-side boundary of the depletion layer, defined as carrier injection, is given as:

$$n_p = \bar{n}_p\left(\exp\frac{-qV_p}{kT} - 1\right) \tag{4.11}$$

where $\bar{n}_p = n_i^2 / N_a$ is the electron concentration in equilibrium. The injection grows exponentially with V_p, as shown in Fig. 4.10. The injected electron concentration gradually decreases with distance from the junction due to recombination with majority carrier holes. Therefore the fixed charge injector should be made close (within a few diffusion lengths) to the MOS capacitor C1.

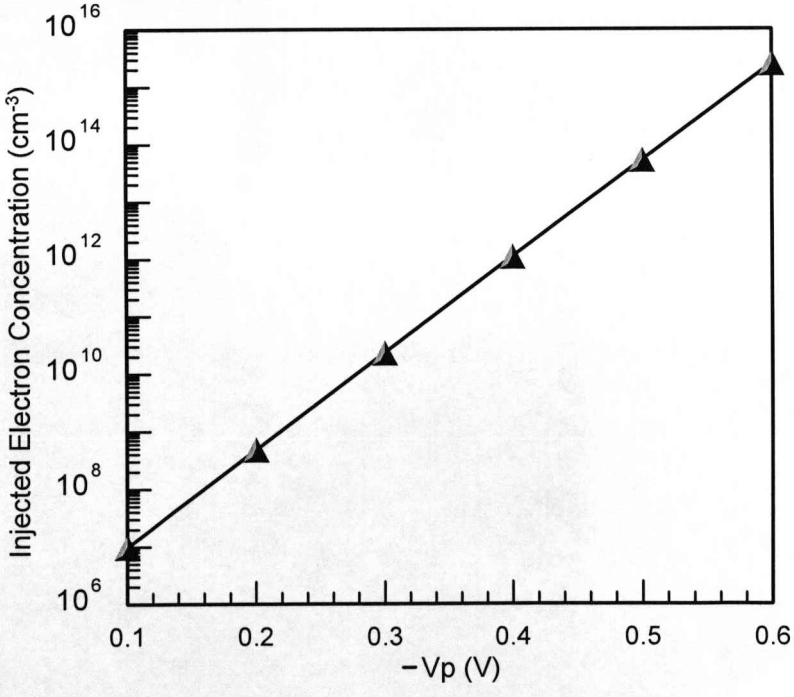


Fig. 4.10 Injected charge concentration for a variety of biasing voltage V_p. N_a=10^{15}cm^{-3}; N_d=10^{19}cm^{-3}.

The simulation results shown in Fig. 4.11 illustrates the stored weight charge Q_W as a function of time where, prior to the application of the voltage spike on C2, Q_W remains unchanged and is approximately $1\times10^{17}cm^{-3}$. When the voltage pulse on C2 is rapidly increased to 2.5V the magnitude of Q_W drops due to the lateral field in the channel between C1 and C2. The 'undershoot' in the charge concentration in the silicon of C1 may be due to the rapid change in the lateral field as the surface potential under the gate of C1 is rapidly modulated.

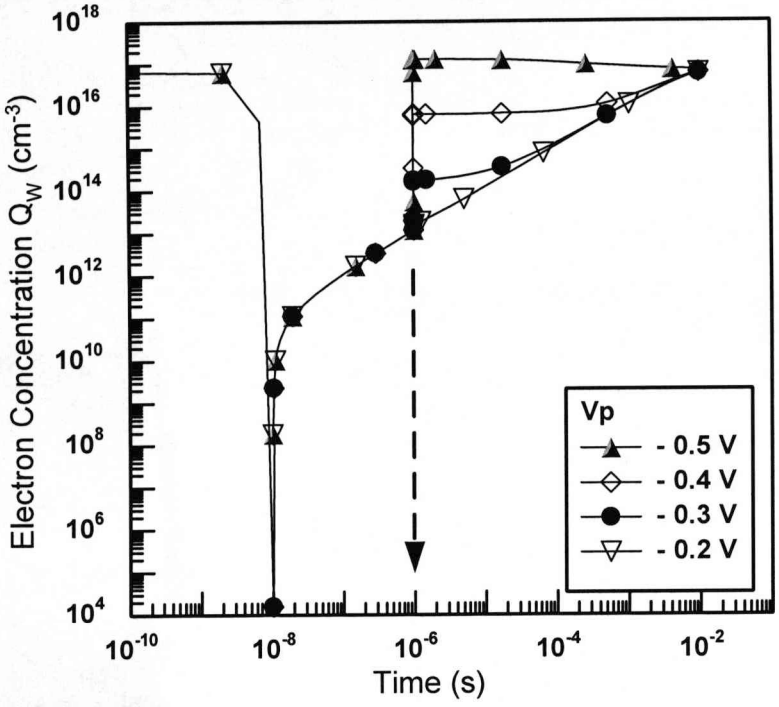


Fig. 4.11: Time-dependent weight charge Q_W under non-equilibrium. t_{ox}=50nm; $N_a=10^{15}cm^{-3}$; $N_d=10^{19}cm^{-3}$; V_{ji}=1V; $\tau_g=10^{-7}$s.

When the voltage pulse at C2 falls after a period of 10ns, the surface potential under the gate of C2 becomes less positive and some of the remaining free charge in C2 is returned to C1 causing a commensurate increase in the minority carrier concentration in C1. Thereafter the charge Q_W accumulates in C1 due to thermal generation of electron-hole pairs. This is a relatively slow process lasting seconds and in the case where the synapse is driven by a spike train whose average frequency is greater than 1Hz, there would be insufficient time between successive spikes to re-establish Q_W.

However, incorporating the minority carrier injector shown in Fig. 4.9, the results of Fig. 4.11 demonstrate that a V_p of -0.5V is sufficient to ensure that Q_W is fully re-established at time t (=1μs) and a V_p of -0.3V enables the re-generation process to be finished within 10ms. It is interesting to note that by varying time, t the operating frequency of the proposed charge coupled synapse is very well defined with a sharp programmable cut-off frequency. Therefore, this structure could implement the dynamic operation associated with biological synapses across a range of cut-off frequencies.

Section 4.5 Programmable Synapse with Injector Transistor

Synapses conduct a signal to a post-synaptic neuron through the excitatory post-synaptic potential (EPSP) or inhibitory post-synaptic potential (IPSP). If the firing frequency of the pre-synaptic neuron is such that the paired pulse ratio, which is the ratio of the amplitude of the second response to that of the first, associated with the EPSP or IPSP response is less than unity, then the synapse is said to be 'depressing' in nature. Conversely, if the ratio is greater than unity, the synapse is said to be 'facilitating'. A single semiconductor device that implements the latter of these transmission states is the focus of this section where a detailed discussion of the operation of the programmable synapse supported by simulation results is now presented.

Section 4.5.1 Device Operation

A schematic of the proposed programmable synapse, consisting of an injector MOS transistor M_i in series with a two capacitor (C1 and C2) structure, is shown in Fig. 4.12. Note that the N+ implants are self-aligned to the polySi contacts: the implants between the electrodes are used to ensure that charge transfers without the need for fringing fields that underpin the charge coupled device (CCD) concept and therefore the programmable synapse can be fabricated using conventional CMOS processing.

This compatibility with standard CMOS represents a major advantage of the structure. The weight charge Q_W is stored in the well under the gate of capacitor C1 where its magnitude is controlled by V_{ji}.

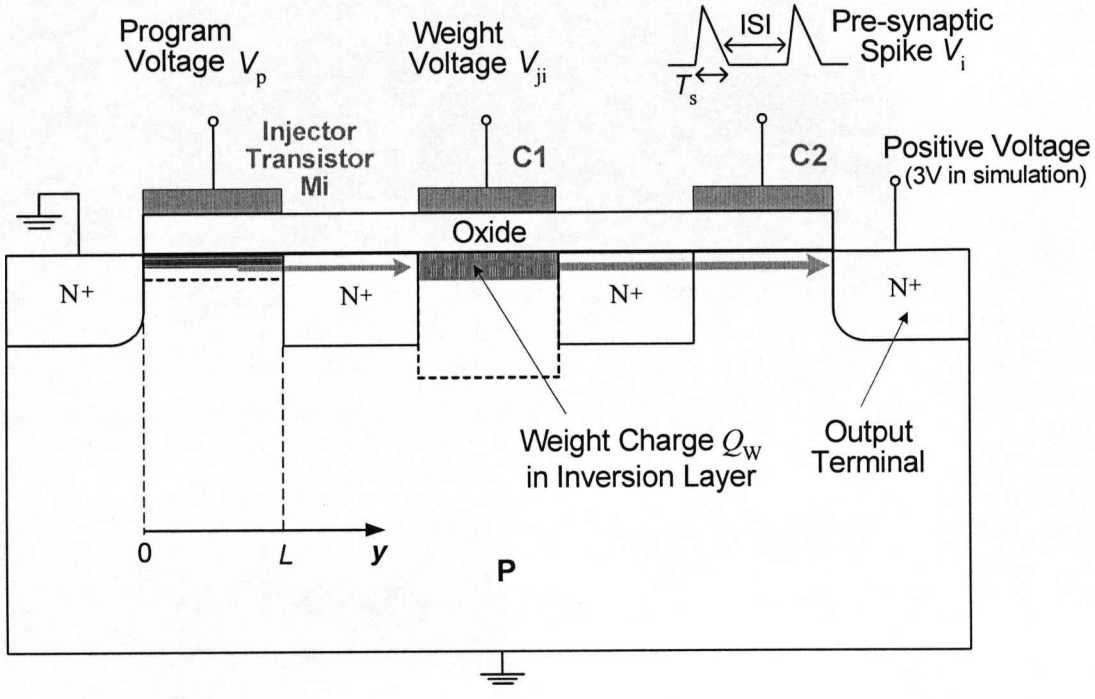


Fig. 4.12 Schematic view of the programmable synapse with an injector transistor M_i.

Consider the case where a spike voltage from a pre-synaptic neuron is applied to the gate of C2 for the duration T_s causing the charge in the well to flow by drift to the output N+ contact. This transfer of charge results in a spike current at the output terminal whose magnitude is dependent on the value of V_{ji}: the charge transfer process can be assumed to be almost 100% efficient and has duration of the order of nanoseconds. During this non-equilibrium condition, the surface potential (at the silicon-oxide interface) in the well of C1 becomes more positive, as the depletion layer widens, and acts as a virtual drain bias for M_i. Therefore, a subthreshold current flows in M_i filling the well with minority carriers. At the instant that the charge density in the well reaches equilibrium, the associated subthreshold current is reduced to zero. Note that according to (2.44), the level of subthreshold current is almost independent of the well voltage (which acts as a virtual drain for M_i) as long as the

effective $V_{ds}>3V_t$. Therefore, the synaptic operation is transient because it depends on the potential induced by the deep depletion condition under the storage well. The magnitude of the subthreshold injector current of M_i is controlled by V_p and consequently the rate at which the well is filled with charge can be controlled. This has implications for frequency encoded data which will be discussed later.

Since the N+ source and the bulk substrate are both grounded, the lateral surface potential is constant and therefore the associated electric field is zero along the channel, as depicted in Fig. 4.12. The MOS transistor theory presented in Chapter 2 is employed to determine the amount of charge injected into the well from the source of M_i. The charge density in M_i is given by (2.40), and the surface potential ϕ_s is related to the gate voltage V_p through (2.8). Assuming that the subthreshold current is independent of the well voltage, the subthreshold current I_{sub} is given by (2.44). The results of this analysis are presented and discussed in the next section.

Section 4.5.2 Analysis and Simulation Results

Simulations on the proposed synaptic device were carried out with Silvaco ATLAS using physical parameters compatible with the AMS 0.35μm CMOS mixed-signal process. The synapse is modeled as three virtual MOS transistors with 0.6μm gate length placed on the $2.12\times10^{17}cm^{-3}$ *p*-type substrate, without the threshold implant. The self-aligned N+ regions are doped to a density of $2\times10^{19}cm^{-3}$ and the electrode spacing is set to 0.6μm. AMS offers a range of gate oxide thicknesses and 16nm is selected in this case, giving a threshold voltage for the synaptic devices (M_i, C1, and C2) of 1V (the flatband voltage V_{FB} is calculated to be -0.986V). The width W and channel length L of M_i are set to 1.8μm and 0.6μm respectively. A positive voltage of 3V is applied to the output terminal. M_i was simulated independently before being integrated into the programmable synapse to verify (2.44). The dependence of the subthreshold current of M_i, at low and high drain bias V_{ds}, on V_p is shown in Fig. 4.13. In the weak inversion region defined by $\phi_B \leq \phi_s \leq 2\phi_B$ (where ϕ_B is the Fermi potential), the subthreshold current has the range $1\times10^{-15}A$ to $1\times10^{-7}A$. For high V_{ds}

there is a parallel shift of the curve due to the lowering of the potential barrier between source and virtual drain, referred to as drain-induced barrier lowering (DIBL). However, DIBL is not considered to be significant because the well voltage under C1 (virtual drain bias) is limited to a few hundred millivolts for the gate voltages on C1 less than 1.8V. As shown in Fig. 4.13, the subthreshold slope starts to degrade for values of V_p lower than 0.2V because conduction in this region is dominated by junction leakage and is process dependent.

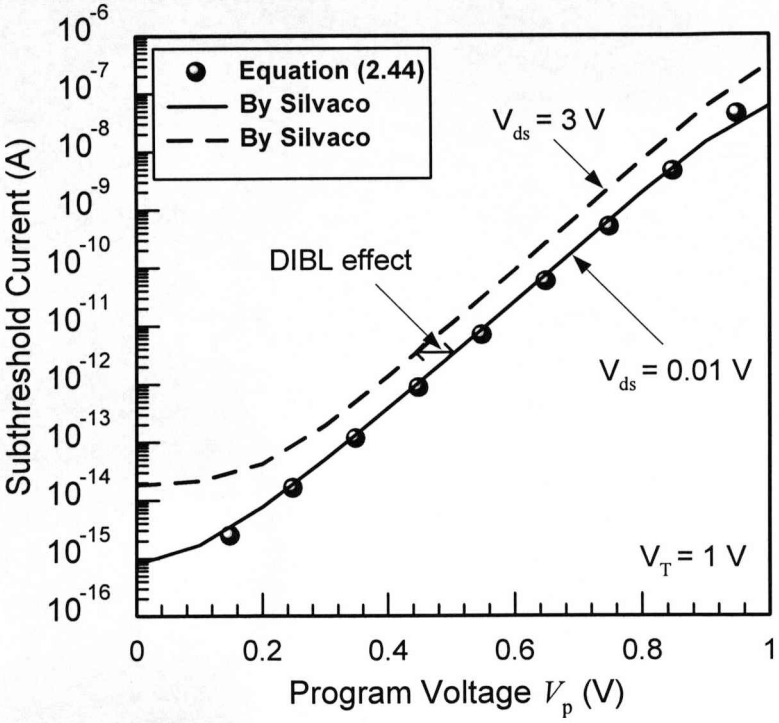


Fig. 4.13 Subthreshold current of M_i as a function of the program voltage V_p showing DIBL. The threshold voltage V_T of M_i is 1V.

As described in section 4.5.1, when a pre-synaptic pulse arrives at the gate of C2, the weight charge Q_W in the well under C1 drifts laterally under the gate of C2 to the output node and the well empties almost instantaneously. However, because M_i is biased in subthreshold, a small current will flow from the N+ source to refill the well: the rate at which the well is filled is determined by V_p. When the well is fully restored, equilibrium is established and no current flows in M_i. The time to establish the equilibrium charge density in the well Q_W as a function of V_p is shown in Fig. 4.14. For this experiment V_p was varied from 0V to 1V in steps of 0.1V and for each value

of V_p a pre-synaptic spike was applied the gate of C2 and the relaxation of the inversion charge in the well was measured. In equilibrium the average electron concentration in the storage well is approximately $10^{18}cm^{-3}$, corresponding to V_{ji} = 2V.

Immediately after the weight charge packet is released by the pre-synaptic spike, the well is empty and the refill duration can be varied from the nanosecond to seconds regime: $V_p = 0$ (thermal generation only) results in a refill time of 1.5s whereas a V_p of 0.5V or 1V gives refill times of approximately 180μs or 10ns respectively. This programmability has major implications for frequency encoded data where, for facilitating behavior, successive pre-synaptic spikes require a paired pulse ratio greater than one. Therefore, the weight storage well must re-establish its equilibrium charge density within the minimum ISI. Referring to Fig. 4.14 and considering the curve for $V_p = 0.5V$, it can be noted that the recovery of 180μs sets the minimum ISI which corresponds to a constant frequency for the pre-synaptic train of approximately 5.5kHz: the minimum ISI is defined as the shortest time to guarantee full recovery of the well charge between successive pre-synaptic spikes.

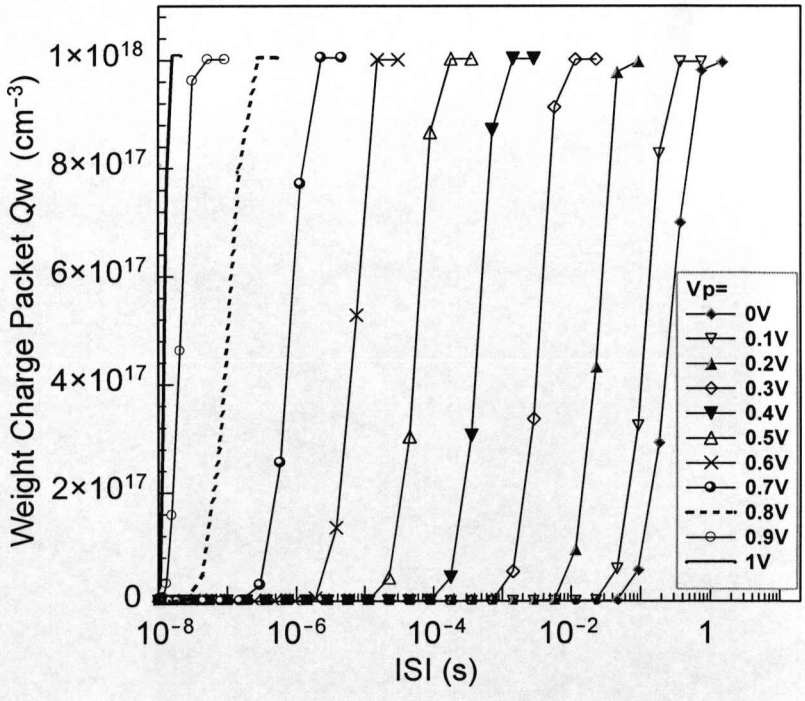


Fig. 4.14 Restoration of the weight charge Q_W as a function of time. V_p is a parameter.

Fig. 4.15 shows the minimum ISI as a function of V_p and for a change in V_p of 1V, the resulting ISI changes over eight orders of magnitude. Therefore, the proposed programmable synapse can not only accommodate information that is encoded as single spikes but also information encoded in spike trains can also be processed: note that for the latter encoding scheme V_p must be set for an ISI corresponding to the maximum instantaneous frequency.

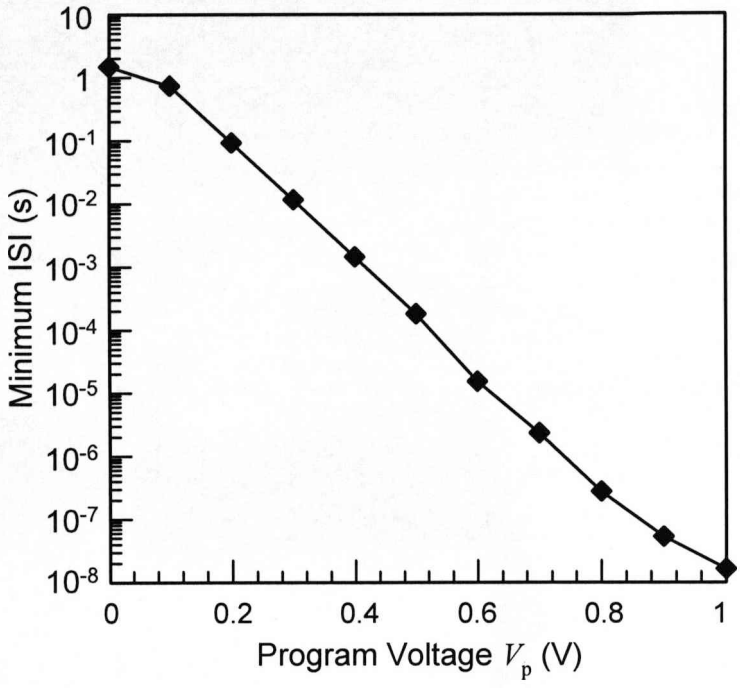


Fig. 4.15 Minimum ISI as a function of program voltage V_p.

Fig. 4.16 shows the amplitude of two successive output spike currents resulting from the application of two successive pre-synaptic spikes: this plot was taken for V_p = 0.3V which corresponds to a minimum ISI of 12ms. The initial output spike, whose detail is shown in the inset, has a magnitude of just over 0.8μA. If the second pre-synaptic spike occurs at the minimum ISI, then the charge in the well is fully re-established and the corresponding output spike should have similar amplitude. Referring to Fig. 4.16 this is indeed the case. However, if the ISI associated with the pre-synaptic spike pair is reduced from the minimum value (12ms), then the second output spike will have smaller amplitude. Fig. 4.16 demonstrates that this is the case where ISIs of 2ms, 4ms and 8ms, have been simulated. Clearly as the ISI is further reduced from its minimum value, the time to re-establish the equilibrium charge

density in the well diminishes and consequently the amplitude of the second output spike is reduced.

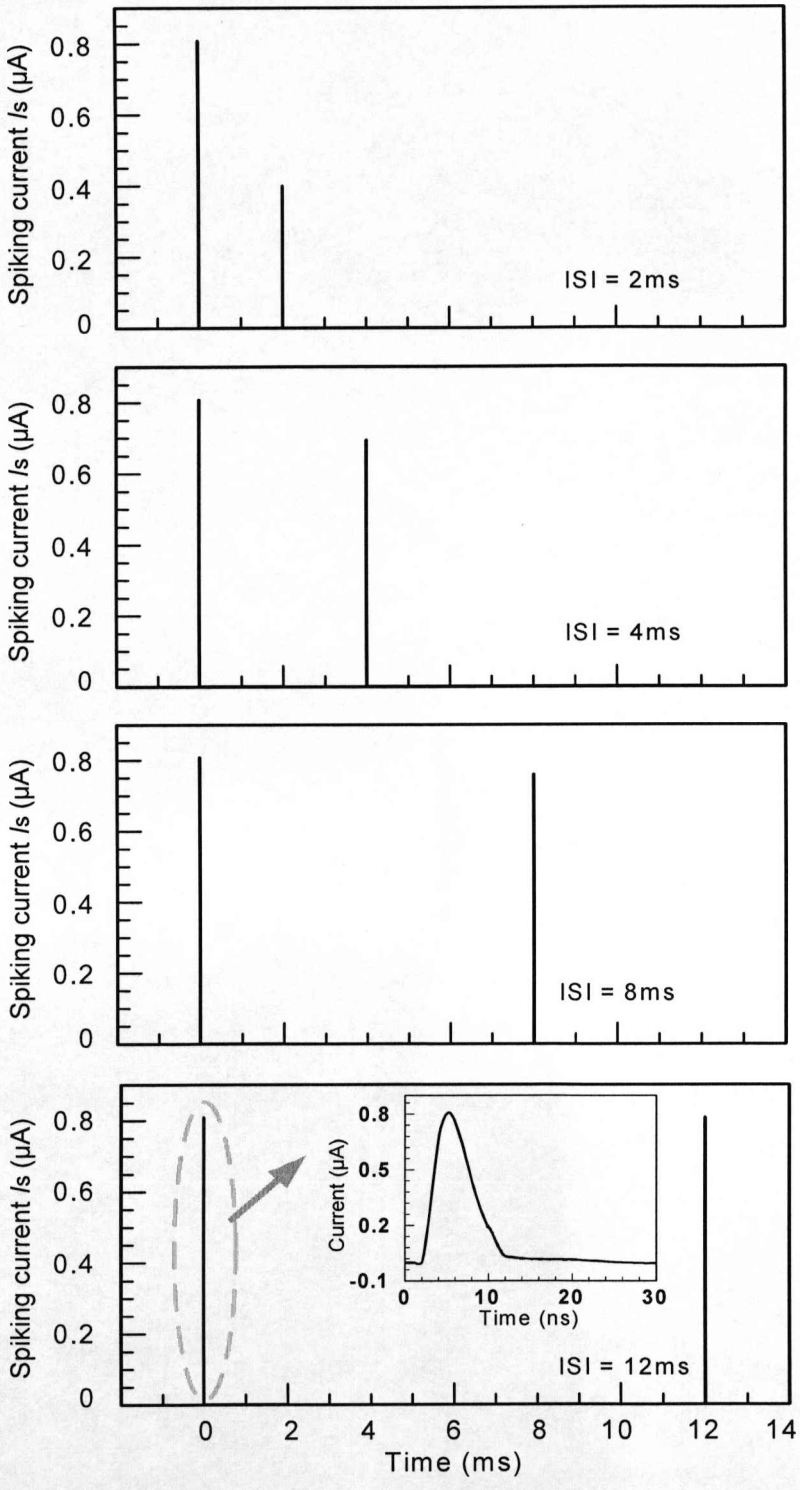


Fig. 4.16 Spiking current at the output terminal of the programmable synapse over time, for ISIs of 2ms, 4ms, 8ms, and 12ms. V_p=0.3V corresponding to a min ISI of 12ms. A single output current spike is shown in the inset.

Section 4.6 Discussion and Conclusions

In this chapter the transient analysis of the charge coupled synapse, proposed in Chapter 3, was presented. The weight charge packet Q_W in the first MOS capacitor C1 is removed by the pre-synaptic pulse applied to the second MOS capacitor C2. After the activation of the synapse the weight charge packet is re-established through the thermal generation of electron-hole pairs. A time-dependent synaptic response is realized at the output terminal. The mathematical analysis and simulation were presented to support this work. It is shown that competition between the natural process of thermal generation in semiconductor junction and ISI can be harnessed to produce either facilitating or depressing synapses.

Since the recovery of C1 from the deep depletion state can take the order of seconds, a minority carrier injector is introduced to control this relaxation process, thereby making the synapse more programmable. A small fixed negative voltage V_p is applied on the fixed charge injector, enabling the forward biasing of the n^+-p junction formed by the N+ implant and the p-substrate of the synapse. The lateral drift of charge onto the output node will result in a transient current spike. The ISI of sequential pre-synaptic signals will influence the spike amplitude due to the time-dependence of the charge generation associated with the recovery of C1. Therefore the pre-synaptic signal can effectively modulate the output spike amplitude according to the fraction of weight charge available for transfer.

This chapter has also presented a programmable dynamic synapse in a single semiconductor device comprising of an injector MOS transistor operating in subthreshold and two MOS capacitors in proximity to the transistor. The injector transistor serves to replenish the charge in the inversion layer of the MOS capacitor, which stores the weight, and consequently provides the synapse with the ability to facilitate over a large frequency range. Simulation results have been presented to show this behavior. A pre-synaptic spike to the second MOS capacitor removes the charge in the inversion layer resulting in a current spike at the output node: this node voltage can be used to stimulate a point neuron circuit. The amplitude of the spike correlates with the level of charge in the well, which is controlled by the associated

gate voltage: this correlation is used to implement synaptic plasticity. The synapse is compact and since it operates in transient mode, its power consumption is negligible.

In conclusion, this chapter presents device concepts for programmable dynamic synapses in spiking neural networks. By using the innate properties of semiconductors, the charge coupled synapse capable of mimicking spiking and programmable dynamics, serves as the compact component for implementing neural networks in silicon. The novel silicon synapses will provide core building blocks that are not only biologically plausible but have the potential to significantly advance the hardware implementation of spiking neural networks towards the biological-scale, using well proven and robust silicon technology.

References

[1] A. M. Zador and L. E. Dobrunz, "Dynamic synapses in the cortex," *Neuron*, vol. 19, no. 1, pp. 1–4, 1997.

[2] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[3] H. Markram and M. Tsodyks, "Redistribution of synaptic efficacy between neocortical pyramidal neurons," *Nature*, vol. 382, pp. 807–810, 1996.

[4] M. V. Tsodkys and H. Markram, "The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability," *Proceedings of the National Academy of Sciences of the USA*, vol. 94, pp. 719–723, 1997.

[5] M. Tsodyks, K Pawelzik, and H. Markram, "Neural networks with dynamic synapses," *Neural Computation*, vol. 10, pp. 821–835, 1998.

[6] R. S. Zucker and W. G. Regehr, "Short-term synaptic plasticity," *Annu. Rev. Physiol.*, vol. 64, pp. 355–405, 2002.

[7] Y. Kanazawa, T. Asai, and Y. Amemiya, "A hardware depressing synapse and its application to contrast-invariant pattern recognition," in *Proc. SICE Annual conference*, Fukui, 2003, pp. 1558–1563.

[8] C. Bartolozzi and G. Indiveri, "Selective attention implemented with dynamic synapses and integrate-and-fire neurons," *Neurocomputing*, vol. 69, pp. 1971–1976, 2006.

[9] F. P. Heiman, "On the determination of minority carriers lifetime from the transient response of a MOS capacitor," *IEEE Tran. Electron Devices*, vol. ED-14, no. 11, pp. 781–784, 1967.

CHAPTER 5 SILICON NEURON STANDARD CELL

Section 5.1 Introduction

In this chapter we first review the operation of a spiking neuron cell containing *n* pre-synaptic neurons $I_1 - I_n$ with associated synapses (S_1–S_n) and one output neuron, I_o, as shown in Fig. 5.1. Consider the case when the pre-synaptic neuron I_1 outputs a spike which eventually reaches synapse S_1. At the instant S_1 receives a spike it must emit a 'weighted' time decaying post-synaptic potential (PSP) where the characteristic shape of this signal is of paramount importance to the computational ability of neuron cells. The PSP is a transient waveform with significantly different rise and fall time constants caused by the loading effect associated with the post-synaptic membrane. At the point neuron, temporal summation of all incoming PSPs is performed and if this aggregate exceeds a threshold, a spike is emitted and communicated along the axon to other neurons: temporal summation gives a measure of the coincidence of all incoming spikes. Therefore, temporal summation of PSPs and thresholding are crucial to information processing in spiking neurons and are also considered in the following hardware model. These biological characteristics provide the basis for this work.

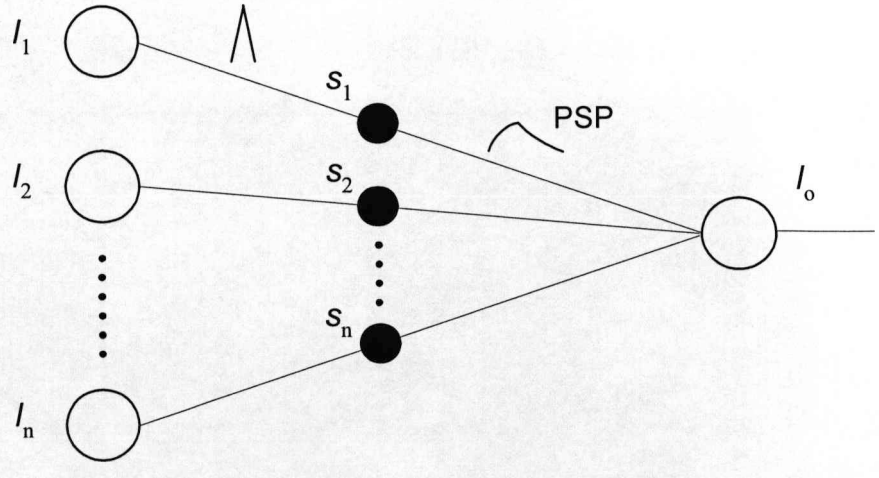


Fig. 5.1 SNN fragment with *n* pre-synaptic neurons, $I_1 - I_n$, and one output neuron, I_o. $S_1 - S_n$ are the synaptic weights.

A large number of neuron circuit ideas have appeared in the literature over the last few decades, as reviewed in Chapter 1. The conductance-based ideas are capable of imitating the processing of the cortex, which is made of a large number of complex non-linear oscillatory neurons exhibiting a variety of inherent firing patterns, but they are intractable and consume large silicon area. The alternative ideas based on the phenomenology are also powerful and can perform firing behavior without complex circuitry. In its simplest form, the integrate-and-fire (I&F) neuron model consists of a capacitor and a threshold device [1]. Fundamentally the neuron cell membrane acts as a capacitor where the potential of the cell membrane can be modeled as the response of a capacitor to an injection of current. In response to a stimulant current the capacitor is charged, and when the potential reaches level, a spike is produced and the potential subsequently returns to the resting potential.

In this chapter, a biologically plausible silicon neuron cell based on the recently developed charge coupled synapses is presented. This silicon synapse produces a weighted spike characteristic using a single device and when embedded in a point neuron cell, a time-dependent output is produced. The signals from charge coupled synapses are integrated as the summed current by a current mirror configuration where the array of synapses can be connected via a common output node. The functionality of thresholding is realized by a CMOS inverter, which accumulates the weighted charges and captures the time-dependency of the post-synaptic membrane decay, mimicked by the charge leakage through a reverse-biased diode or a leaky transistor. Correspondence is made between the semiconductor relaxation processes and biologically relevant responses such as PSP and refractory period.

The rest of the chapter is organized as follows. In section 5.2, the principles and operation of the neuron cell circuit are discussed. Section 5.3 presents the simulation results of the neuron cell circuit. The simulation results on the integration of programmable synapses described in Chapter 4 are presented and discussed in Section 5.4 and Section 5.5. Section 5.6 deals with a description of an alternative silicon neuron cell based on the conventional neuMOS device. Discussion and conclusions are given in section 5.7.

Section 5.2 Analog Neuron Cell Circuit

In ideal spiking neurons if the sum of the inputs, from different dendrites, surpasses a particular threshold then a spike is produced which propagates along the axon to other synapses. A simple implementation of I&F model is the axon-hillock circuit proposed in [2], which comprises an integrating capacitor connected to two inverters, a feedback capacitor, and a reset transistor driven by the output inverter. They are widely used in the realization of neural coding and large networks. The similar concept presented here however, represents a new paradigm to incorporate a spiking neuron with the novel charge coupled synapses, which has the potential to build more biologically plausible neural networks in hardware towards compactness and low power.

The circuit diagram of the proposed silicon neuron cell is shown in Fig. 5.2. M1-M2 constitutes a current mirror which is used to facilitate the integration of the weighted current spikes from a number of *n*-type charge coupled synapses. Current mirror action reflects $i(t)$ at the drain terminal of M2 and we define the time-dependent voltage at this terminal as V_{PSP}: the post-synaptic membrane node potential. This current charges the membrane node which drives the CMOS inverter consisting of M3 and M4, hereafter defined as $CMOS_{3,4}$, while simultaneously a reverse-biased diode D1 provides a leakage path for the discharge of this node. This charge leakage effectively mimics the decay of the membrane potential of biological neurons (Note that a MOS transistor biased in subthreshold could also be employed as a leakage path and this could offer the ability to tune the V_{PSP}). A sufficiently large number of synaptic outputs will cause V_{PSP} to reach the switching threshold of $CMOS_{3,4}$, which will generates an output hi-lo transition. As a result, the output of $CMOS_{5,6}$, which is fed to the subsequent synapses, makes a lo-hi transition, thus turning on M7. The neuron is said to have fired. The gate of $CMOS_{3,4}$ is discharged through M7 and V_{PSP} is reset to zero until the time when another synaptic input is received.

The charge coupled synapses presented in Chapter 3 are directly connected to the drain of M1 via a common output terminal. The leakage of charge through the n^+-p junction at the synapse output terminal and substrate, is controlled by thermal generation and this process is orders of magnitude greater than the aforementioned

charge transfer time. Therefore we can consider that no charge is lost during the summing operation. A refractory period is present after the integration of the spiking currents because the recovery of the silicon synapses from the deep depletion conditions takes the order of milliseconds when the lifetime quenching is employed. Note that the programmable dynamic synapses presented in Chapter 4, are capable of injecting minority carrier electrons to facilitate and control this relaxation process.

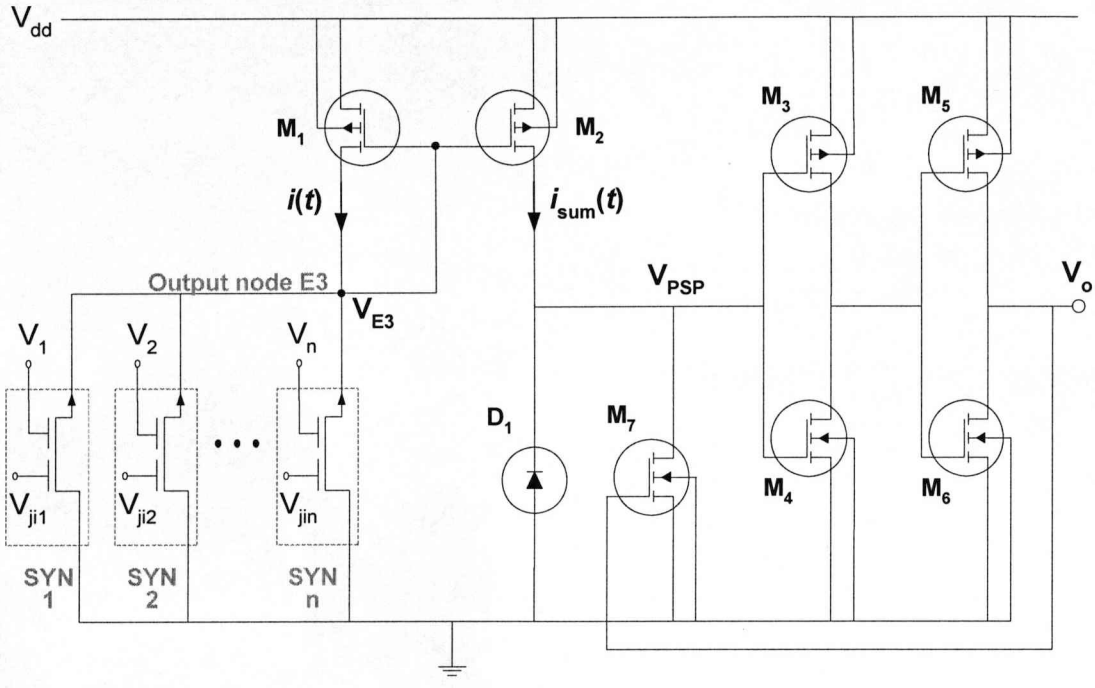


Fig. 5.2 The analog spiking neuron circuit with an array of *n*-type charge coupled synapses described in Chapter 3. The integration and thresholding functions are implemented by using the current mirror configuration (M1 and M2) and CMOS inverters ($CMOS_{3,4}$; $CMOS_{5,6}$) respectively. Leaky diode D1 provides a leakage path for the discharge of the membrane node. The feedback to M7 ensures that the neuron resets after firing.

Section 5.2.1 Synapses-Neuron Interfacing

Firstly, consider the charging/discharging voltage on the output terminal of the synapses, V_{E3}, when the weight charge packets Q_W are being transferred from the array of synapses to the input of $CMOS_{3,4}$. The node voltage transient, δV_{E3}, will be determined by the node capacitance and to estimate this capacitance, assume that there are n synapses common to this node. Therefore the associated capacitance is

nC_{FN}, where C_{FN}, defined in Fig. 3.6, is the depletion capacitance associated with the n^+-p junction at the output of the charge coupled synapse. The total capacitance of the common output node is therefore $2C_{gs}//nC_{FN} \sim nC_{FN}$ for large n: C_{gs} is the gate-source capacitance of M1 and M2, hence the factor 2. Assuming that the charge transfer from the weight storage well to output terminal is instantaneous, then the output node voltage will drop by δV_{E3} and subsequently increase because of the charging current through M1. Consider a worst case where δV_{E3} is such that M1 is operating in the subthreshold region for the duration of the transfer with a drain current, $i(t)$, given by:

$$i(t) = I_0 \exp\left(\frac{\delta V_{E3}}{mV_t}\right) \tag{5.1}$$

where V_t is the thermal voltage; m and I_0 are the gate channel coupling and off-current associated with M1 respectively: $m = 1.5$, equivalent to a typical subthreshold slope of 90mV/decade, and $I_0 \sim 10^{-12}$A are typical values consistent with a threshold voltage of about 0.7V. To obtain an expression for the time-dependency of δV_{E3} we can write that:

$$I_0 \exp\left(\frac{\delta V_{E3}}{mV_t}\right) = nC_{FN}\frac{d\delta V_{E3}}{dt} \tag{5.2}$$

Solving (5.2) with the initial condition $\delta V_{E3} = \delta V_{E3}$ ($\sim Q_W/nC_{FN}$) at $t = 0$, and final condition $\delta V_{E3} = 0$ at $t = \tau_r$, gives:

$$\frac{I_0\tau_r}{nC_{FN}mV_t} = 1 - \exp\left(\frac{-\delta V_{E3}}{mV_t}\right) \tag{5.3}$$

where τ_r is the duration of the transfer of Q_W from synaptic output node to the input of $CMOS_{3,4}$: essentially τ_r is the time taken for V_{PSP} to reach its peak value, which sets the fan-in for the neuron circuit. For biological neurons, τ_r is typically of the order of a few milliseconds and for the purpose of the following calculation a value for τ_r of 5ms is used. Based on the approximation that $\delta V_{E3} > mV_t$ for most of the charge transfer process, (5.3) can be re-arranged to give an estimate of the fan-in, n, for the specified τ_r:

$$n \approx \frac{I_0\tau_r}{C_{FN}mV_t} \tag{5.4}$$

Consider a 1μm process where the substrate doping is 10^{15}cm^{-3}, and assume a one-sided step junction approximation, C_{FN} is estimated to be of the order of 5×10^{-16}F.

Therefore, substituting a value for τ_r of 5ms into (5.4) gives an approximate value for n of 266. The analysis highlights that the maximum number of synapses in an array that can share the output node is limited by τ_r and other process dependent parameters. However, the number of synaptic inputs can still be scaled by having k arrays where the total number of synapses (fan-in) would then be $n \times k$ in a massively parallel computing system.

Section 5.2.2 Current Mirror Operation

All n synaptic output signals are integrated onto the drain of M1. The transfer of weight charge packets results in a current $i(t)$. Two p-channel devices which can be triggered by $i(t)$ are required in the current mirror configuration. Therefore the accumulated charge is transferred to the drain of M1, the reference current $i(t)$ increases and is mirrored as $i_{sum}(t)$ in M2. The current $i_{sum}(t)$ charges the node and hence gate voltage V_{PSP} of the first CMOS inverter, and D1 is then under reverse bias.

Assume matched p-channel MOS transistors, M1 and M2. Consider the case that the signal from the synaptic array is large enough to enable M1 and M2 to operate above threshold. The current for M1 (assuming a long channel device) in saturation, whose source and substrate is connected to V_{dd}, is given by:

$$i(t) = \mu C_{ox} \frac{W_1}{L_1} \frac{(\delta V_{E3}(t) - V_T)^2}{2m} \tag{5.5}$$

where μ is the mobility; C_{ox} is the oxide capacitance; W_1 and L_1 are the channel width and length of M1 respectively; m is the gate channel coupling.

By re-arranging (5.5), the current $i(t)$ from the synaptic array sets the V_{E3} to:

$$V_{E3}(t) = V_T + V_{dd} - \sqrt{\frac{2i(t)L_1}{\mu C_{ox} W_1}} \tag{5.6}$$

The transistor M2 is then controlled by V_{E3}, and the saturation current $i_{sum}(t)$ is obtained:

$$i_{sum}(t) = \mu C_{ox} \frac{W_2}{L_2} \frac{(\delta V_{E3}(t) - V_T)^2}{2m} \tag{5.7}$$

where W_2 and L_2 are the channel width and length of M2 respectively; m is the gate channel coupling.

Substitute (5.6) into (5.7), the current $i_{sum}(t)$ is expressed as a function of $i(t)$:

$$i_{sum}(t) = \alpha \times i(t) \tag{5.8}$$

where $\alpha = W_2L_1/(W_1L_2)$ is the ratio of aspect ratios for M1 and M2. Note that $\alpha > 1$ allows current signals to have a fan-out greater than one and each output can be scaled using an appropriate W/L ratio. If M1 and M2 are matched in all respects, $i_{sum}(t) = i(t)$. After the spike emission period, V_{E3} is driven back to V_{dd} causing the resetting of the common output terminal of the synaptic array.

In practice, there will be a mismatch error of the current mirror due to variation in the W/L ratios and threshold voltages. The geometry-dependent mismatch contributes a fractional current error that is independent of bias, and the mismatch error due to threshold voltages increases as $(\delta V_{E3} - V_T)$ is reduced. The current mirror may also generate an error, arising from the mismatch of drain voltages as the current depends on the drain-source voltage for each transistor. This 'early effect' is slight for MOS transistors acting in subthreshold. Above threshold, the output characteristic is not very flat unless we make the channel lengths (L_1 and L_2) large. However, these mismatch errors are small compared with the total output current from the synaptic array, so can be negligible in the implementation.

Section 5.2.3 Membrane Potential Generation

Referring to Fig. 5.2, consider n parallel connected synapses each of which generates a weighted transient current spike where the total current from all spikes $i(t)$ is summed at the drain terminal of M1 as:

$$i(t) = \sum_{0}^{n} i_n(V_{ji,n}) \tag{5.9}$$

where $i_n(V_{ji,n})$ is the weighted current spike associated with the nth synapse; $V_{ji,n}$ is the associated weight voltage. If ideal mirror operation is assumed, then $i(t)$ will flow in the drain terminal of M2 charging the gates of $CMOS_{3,4}$ and consequently V_{PSP} will increase. To obtain a quantitative analysis of the discharging of the membrane voltage

V_{PSP} consider the case where only the nth synapse has activated and the time taken to transfer the associated charge to the gate of $CMOS_{3,4}$ is τ_r. In this case the maximum membrane voltage value at $t = \tau_r$ is defined as:

$$V_{PSP}(\tau_r) = \frac{1}{C_{ON}(V_{PSP}(\tau_r))} \int_0^{\tau_r} i_n(V_{ji,n}) dt \tag{5.10}$$

where $C_{ON}(V_{PSP}(\tau_r))$ is its associated membrane node capacitance. The leakage path is via the reverse-biased diode D1 as shown in Fig. 5.2. In order to analyze this discharge, the equivalent circuit of Fig. 5.3 is considered. $C_D(V_{PSP})$ represents the depletion capacitance of the diode D1, and C_p the capacitance at that node, namely the input capacitance of $CMOS_{3,4}$, the drain capacitance of M2 and other parasitic capacitance. It is assumed that the capacitances represented by C_p are independent of voltage for the purpose of this analysis.

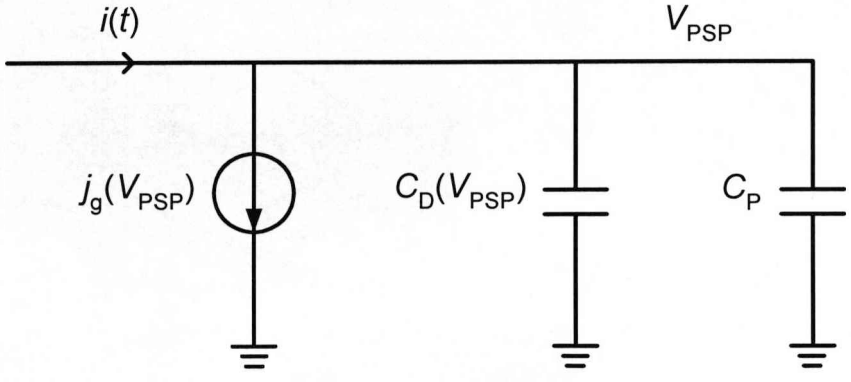


Fig. 5.3: Equivalent circuit for the discharging of the membrane node of the neuron circuit.

The charge on the membrane node capacitance $C_{ON}(V_{PSP})$ (= $C_D(V_{PSP}) + C_p$) will leak away through the reverse-biased depletion region associated with the diode according to:

$$j_g(V_{PSP}) = -\left[C_D(V_{PSP}) + C_p\right]\frac{dV_{PSP}}{dt} \tag{5.11}$$

Assuming a one-sided step junction diode we have that:

$$C_D(V_{PSP}) = \frac{\varepsilon_{si}\varepsilon_0}{W_d(V_{PSP})} \tag{5.12}$$

where $W_d(V_{PSP})$ is the voltage (and hence time) dependent depletion width associated with $C_D(V_{PSP})$. The leakage current $j_g(V_{PSP})$ in silicon is dominated by thermal

generation of electron-hole pairs in the reverse-biased depletion region associated with the diode. Hence it can be obtained that:

$$j_g(V_{PSP}) = \frac{qn_i W_d(V_{PSP})}{\tau_g} \tag{5.13}$$

where τ_g is the minority carrier generation lifetime.

Equating (5.11) and (5.13), substituting for $C_D(V_{PSP})$ using (5.12) and separating variables gives:

$$\frac{n_i}{\tau_g}\int_{\tau_r}^{t} dt = \frac{1}{q}\int_{V_{PSP}(\tau_r)}^{V_{PSP}} \left[\frac{\varepsilon_{si}\varepsilon_0}{W_d(V_{PSP})^2} + \frac{C_p}{W_d(V_{PSP})}\right] dV_{PSP} \tag{5.14}$$

where $V_{PSP}(\tau_r)$ is the spike voltage given by (5.10). From one-sided *p-n* junction theory the depletion width is expressed as:

$$W_d(V_{PSP}) = \sqrt{\frac{2\varepsilon_{si}\varepsilon_0(V_{bi} + V_{PSP})}{qN_a}} \tag{5.15}$$

where V_{bi} is the built-in voltage of the *p-n* junction. Substituting for $W_d(V_{PSP})$ in (5.14) using (5.15) and integrating gives:

$$t(V_{PSP}) = \frac{\tau_g N_a}{2n_i}\left[\ln\left(\frac{V_{bi} + V_{PSP}}{V_{bi} + V_{PSP}(\tau_r)}\right) + C_p\sqrt{\frac{8(V_{bi} + V_{PSP}(\tau_r))}{qN_a\varepsilon_{si}\varepsilon_0}}\left(\sqrt{\frac{V_{bi} + V_{PSP}}{V_{bi} + V_{PSP}(\tau_r)}} - 1\right)\right] \tag{5.16}$$

which is a transcendental equation in V_{PSP} and is readily solved numerically. It should be noted that the approximation that $\tau_r << t(V_{PSP})$ has been made. If we consider n spikes then each spike will add a voltage increment, whose magnitude is given by (5.10), to the input summing node of $CMOS_{3,4}$. However, the maximum voltage at this node will depend on the weighting of each of the current spikes and their temporal distribution. Given that enough current spikes arrive within the time frame for temporal summation to cause the voltage at $CMOS_{3,4}$ input to equal its switching threshold, then the inverter will change state and the neuron is said to have fired. Note that because the output of the neuron will drive charge coupled synapses on the next layer, it is only required to generate a voltage pulse lasting a few nanoseconds which is sufficient time to release the weight charge packet Q_W, as described in Section 3.4. With reference to Fig. 5.2 the feedback path (V_o to M7) ensures that the neuron cell is reset after firing. Prior to the firing state the output of the neuron cell V_o is 0V and when the cell fires V_o increases rapidly causing the gate voltage on M7 to increase. This action discharges the membrane node, the neuron resets and the output V_o is

constrained by the feedback to be a very short voltage spike. However, at architectural level this simple feedback arrangement will need to be modified to incorporate a process independent delay which will ensure that the output of the second inverter has a fan-out capable of driving multiple subsequent synapses.

Section 5.2.4 Thresholding Operation

The CMOS inverters are employed in the circuit to perform the thresholding function which plays an important role in neuron computation. The switching threshold V_{Th} for CMOS inverter is defined at the point where the input and output voltages are equal. Assume the voltage supply is high enough so that the transistors operate in saturation. The expression for V_{Th} is given by:

$$V_{Th} = \frac{\left(V_{T4} - \frac{V_{sat4}}{2}\right) + r\left(V_{dd} + V_{T3} + \frac{V_{sat3}}{2}\right)}{1+r} \tag{5.17}$$

where V_{T3} and V_{T4} are the threshold voltage of M3 and M4; V_{sat3} and V_{sat4} are the drain saturation voltages for M3 and M4 respectively; $r=W_3L_4V_{\mathrm{sat3}}/(W_4L_3V_{\mathrm{sat4}})$ assuming equal oxide thickness. Therefore the switching threshold of the electronic neuron is determined by appropriate sizing of M3 and M4 although it is worth noting that application of substrate bias to one or both of those transistors allows further adjustment of the threshold via V_{T}. We do not discuss further the thresholding function here since it is the typical operation of a CMOS inverter.

Section 5.3 Simulation Study of Neuron Cell Circuit

There are mainly two kinds of synapses: excitatory synapse and inhibitory synapse. Excitation often results in synchrony and inhibition in asynchrony. Recent studies have revealed that mutual inhibition impedes spiking, pushing synchronous neurons apart, whereas mutual excitation brings the neurons together [3]. However, these relationships can be reversed by synaptic delays. When inhibition lags network activity, it pushes out-of-phase neurons into phase in subsequent cycles, promoting synchrony. Intuitively, delay provides an opportune period for neurons to spike

together before inhibition arrives. On the contrary, delayed excitation impedes synchrony by promoting out-of-phase spiking [4]. We now examine the synchrony and asynchrony of the neuron circuit triggered by the excitatory charge coupled synapses. The analog neuron circuit is simulated in PSpice. The models are those from an AMS 0.35μm CMOS mixed-signal process and are summarized in Table 5.1. The threshold voltages of the *p*- and *n*-channel transistors are -0.68V and 0.49V respectively. For all the transistors, the same width and length are used (W=L=1.2μm). The output of the charge coupled synapses is represented by using current pluses of amplitude 1.2μA for each activated synapse. The power supply V_{dd} was set to 3V.

Table 5.1 0.35μm Spice model parameters

Parameter	PMOS	NMOS
W	1.2 μm	1.2 μm
L	1.2 μm	1.2 μm
LEVEL	2	2
LD	0.15 μm	0.15 μm
TOX	7.754E-9 m	7.575E-9 m
VTO	-0.68 V	0.49V
KP	6.634E-5 A/V^2	2.161E-4 A/V^2
NSUB	101E+15 cm^{-3}	212E+15 cm^{-3}
GAMMA	0.4 $V^{1/2}$	0.58 $V^{1/2}$
PHI	0.815 V	0.853 V
UO	148 cm^2/V-s	475.8 cm^2/V-s
UEXP	0.324	0.324
UCRIT	1.854E+7 V/cm	2.125E+9 V/cm
DELTA	0.01	1.442E-2
VMAX	1.158E5 m/s	1.338E5 m/s
XJ	3E-7 m	3E-7 m
LAMBDA	0.06 V^{-1}	0.06 V^{-1}
NFS	1E+12 cm^{-2}	1E+12 cm^{-2}
NEFF	0.585	2.541
NSS	1E+11 cm^{-2}	1E+11 cm^{-2}
RSH	129 Ω/sq.	70 Ω/sq.
PB	1.02 V	0.69 V
CGDO	8.6E-11 F/m	1.2E-10 F/m
CGSO	8.6E-11 F/m	1.2E-10 F/m
CJ	1.36E-3 F/m^2	9.4E-4 F/m^2
MJ	0.56	0.34
CJSW	3.2E-10 F/m	2.5E-10 F/m
MJSW	0.43	0.23

Section 5.3.1 Synchronous Signal Response

In this experiment, 100 synapses were activated concurrently and the accumulated peak in current is 120μA. The membrane node voltage V_{PSP} at the gates of $CMOS_{3,4}$ is shown in Fig. 5.4. The voltage, following the rise of the synaptic signal, has a fast rise time reaching a peak value 2.9V which causes the $CMOS_{3,4}$ inverter to change state, causing a state transition of $CMOS_{5,6}$. This demonstrates the firing of the biological neuron. Due to the charge leakage through the reverse-biased D1, the membrane potential shown has a much slower fall time of the order of milliseconds, exhibiting the characteristic shape observed in biological neurons. However, the leakage current is temperature dependent, and as previously stated, there is a need to increase the rate of the decay to take advantage of the much higher speed of electronics compared to biological processes. This will be addressed later.

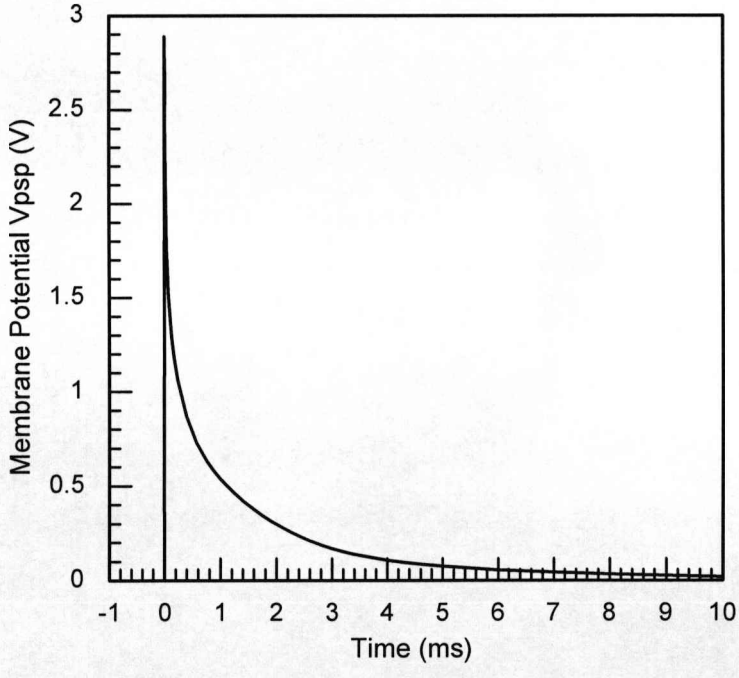


Fig. 5.4 The membrane potential V_{PSP} when the spikes are generated by the synapses at the same time.

Section 5.3.2 Asynchronous Signal Response

Fig. 5.5(a) shows the accumulation of successive PSP signals when the synapses emit spikes with various synaptic time lags. In this simulation, there are 30 active synapses

divided into 6 groups where each synapse is activated at the same time with an interval between each group. The feedback transistor M7 was removed in this instance to prevent the neuron cell from resetting. The associated accumulation of charge produces the V_{PSP} shown in Fig. 5.5(a). Therefore, the threshold is reached and the $CMOS_{5,6}$ inverter is triggered to send out a spike, as shown in Fig. 5.5(b). The plots demonstrate the ability of neuron circuit to aggregate successive input signals over time, in a manner reflective of biological neurons. Because the membrane node discharging is of the order milliseconds, the accumulation performance is limited by the interval between each group. If we assume 3ms interval, the majority of charges are used to compensate the leakage rather than commit to neuron's firing.

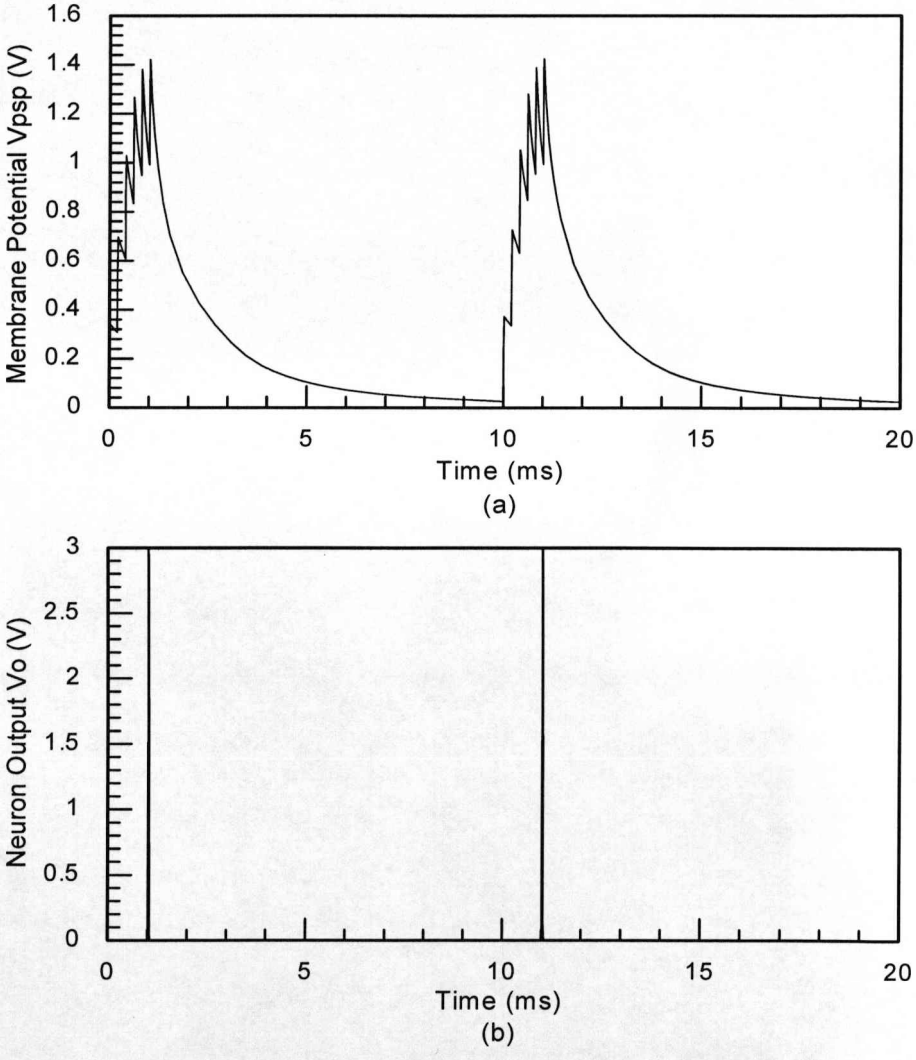


Fig. 5.5 (a) The membrane potential V_{PSP} when a series of spikes are generated with various time lags; (b) The output voltage of the second inverter $CMOS_{5,6}$. The neuron fires when the membrane potential exceeds the threshold of CMOS inverter.

The ability of the neuron circuit to reset itself after firing is demonstrated in Fig. 5.6. When the switching threshold of $CMOS_{5,6}$ is reached, it undergoes a lo-hi transition and the neuron is said to have fired, at which point M7 is turned on and V_{PSP} is reset to 0V. Any spike arriving after the resetting of the membrane node will contribute to the subsequent firing events. Both the switching threshold of the inverters and the duration of V_o are set by the aspect ratios of the transistors in the inverter array; further adjustments would be possible through the application of a substrate bias.

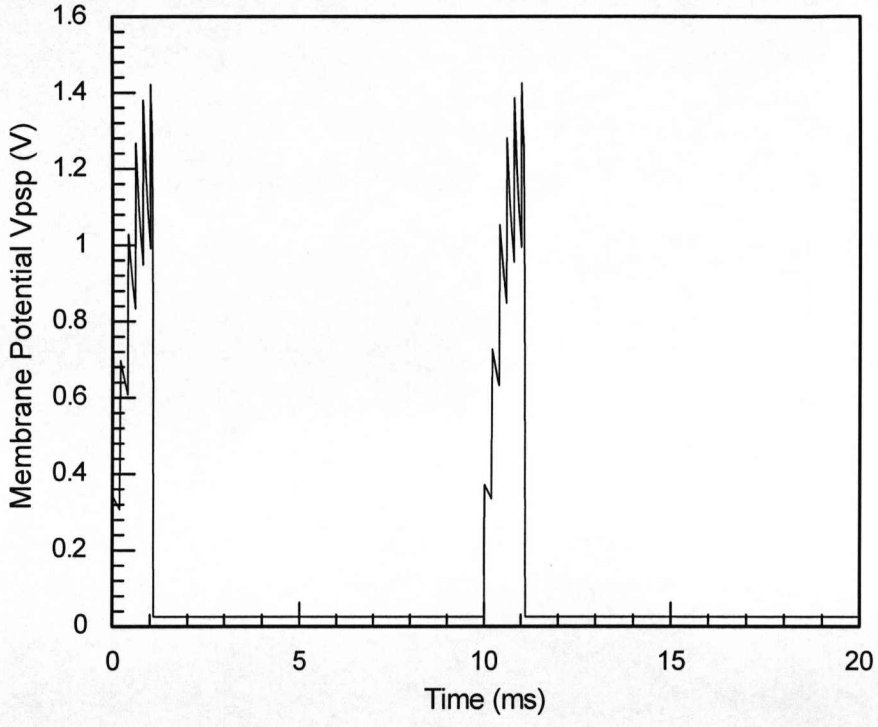


Fig. 5.6 The membrane potential V_{PSP} is reset after neuron's firing.

An estimate of the time required for V_{PSP} to decay away for a given V_{PSP}, as expressed by (5.16), is plotted in Fig. 5.7. From the shape of the plots obtained with the lifetimes of 1μs and 0.1μs, it can be concluded that for smaller spikes the decay time of V_{PSP} is proportional to $\sqrt{V_{PSP}}$, while for larger spikes the decay time can be expressed as a function of $\ln(V_{PSP})$, which is more characteristic of biological neurons. The result shows clearly that biologically plausible membrane potentials are realizable for a

range of spike voltages. It should be noted that τ_g~1-10μs or greater is not untypical for production grade silicon, resulting in rather long relaxation times in the range of ten to hundreds of milliseconds. However, there are reliable and reproducible methods, typically used in power devices, to 'quench' the lifetime. By such means we can reduce τ_g to the nanosecond regime and hence $t(V_{PSP})$ to biologically plausible values of a few milliseconds apparent in Fig. 5.8. As an alternative, 'leakier' diodes with reverse leakage dominated by the Zener tunneling process, which is weakly dependent on temperature, could be employed, as described in the next section. Such a diode could be realized using the source implants for the *p*- and *n*-channel devices within a CMOS process flow. By these means, biologically plausible PSP response could also be realized.

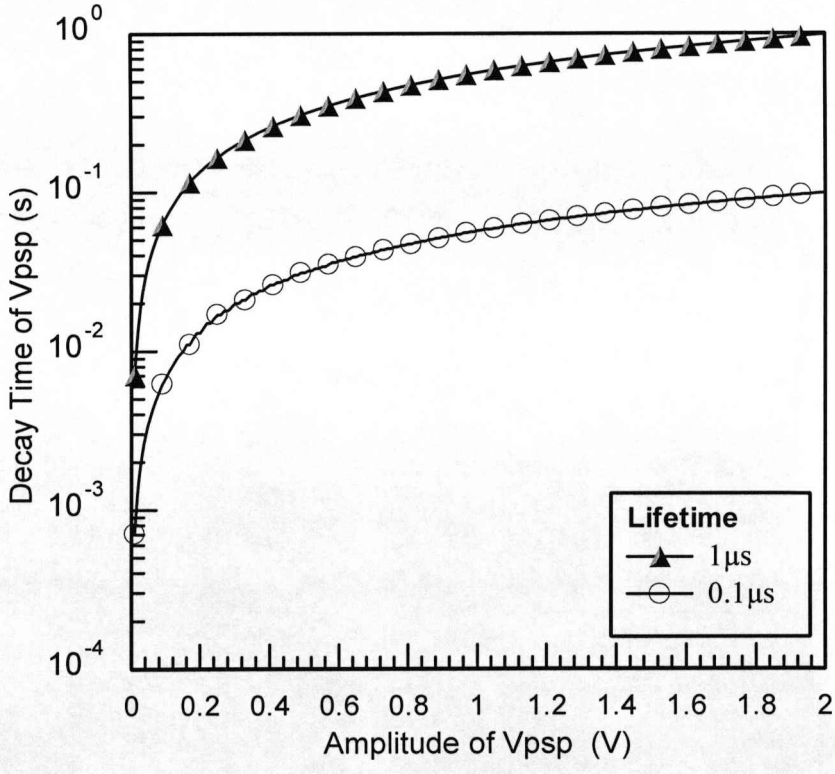


Fig. 5.7 Total decay time of V_{PSP} as a function of the amplitude of V_{PSP} with generation lifetime as a parameter. Curves are for τ_g=1μs and 0.1μs.

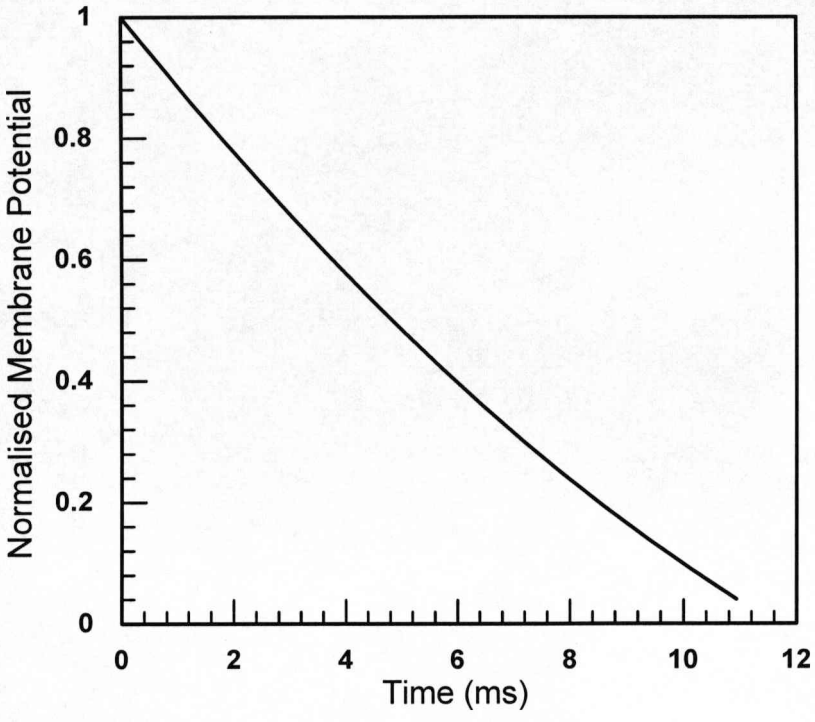


Fig. 5.8 V_{PSP} (normalized) decay on the membrane node (input node of $CMOS_{3,4}$) for τ_g=0.01μs, as predicted by (5.16).

Section 5.4 Integration of Synapses with Charge Injector

As described in Chapter 4, the charge coupled synapse requires times of the order of seconds to re-establish the weight charge packet Q_W after the activation of the synapse. A minority carrier injector, presented in Section 4.4, is introduced to control this relaxation process. This serves to make the synapse more programmable and less process dependent. The programmable synapse can implement operation associated with biological synapses by setting the ISIs to microseconds. The neuron cell contains many synapses whose outputs are aggregated to produce a membrane voltage response as illustrated in Fig. 5.9. Therefore if the program voltage V_p of the synapse is set to -0.3V, and the pre-synaptic spike frequency is restricted to 100Hz, the facilitation behavior of synapses and neurons can be implemented. If the spike frequency goes beyond 100Hz, the depression behavior will be observed. The circuit was simulated in PSpice using 0.35μm CMOS technology parameters given in Table

5.1. V_{dd} was set to 3V. Input current pluses were presented to the circuit representing the dynamics of the outputs from the synapses. The transistor for the resetting of the membrane node was removed in the simulation here.

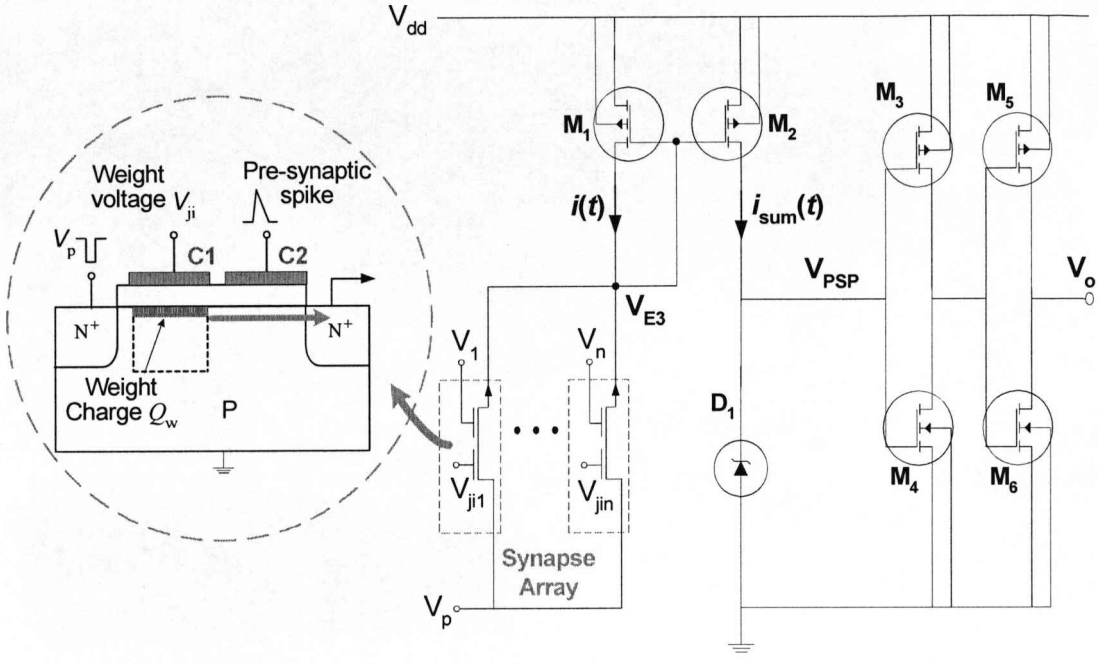


Fig. 5.9 Neuron cell circuit with its associated array of charge coupled synapses with fixed charge injectors.

A low breakdown (4.7V) leaky zener diode D1N750, whose reverse leakage is dominated by Zener tunneling, is employed to form the leakage path (diode D1 in Fig. 5.9). This diode allows for the discharge of the V_{PSP} which in turn, mimics the decaying of the membrane potential in biological neurons. Fig. 5.10 shows the reverse I-V curve of the zener diode. Good correlation between the modeled and measured characteristic is evident. A 'leaky diode' is readily formed using the contact implants of *p*-type and *n*-type MOS transistors. It has the advantages of very small temperature dependence, as Zener dominated leakage depends on the temperature dependence of the energy band gap, and compactness.

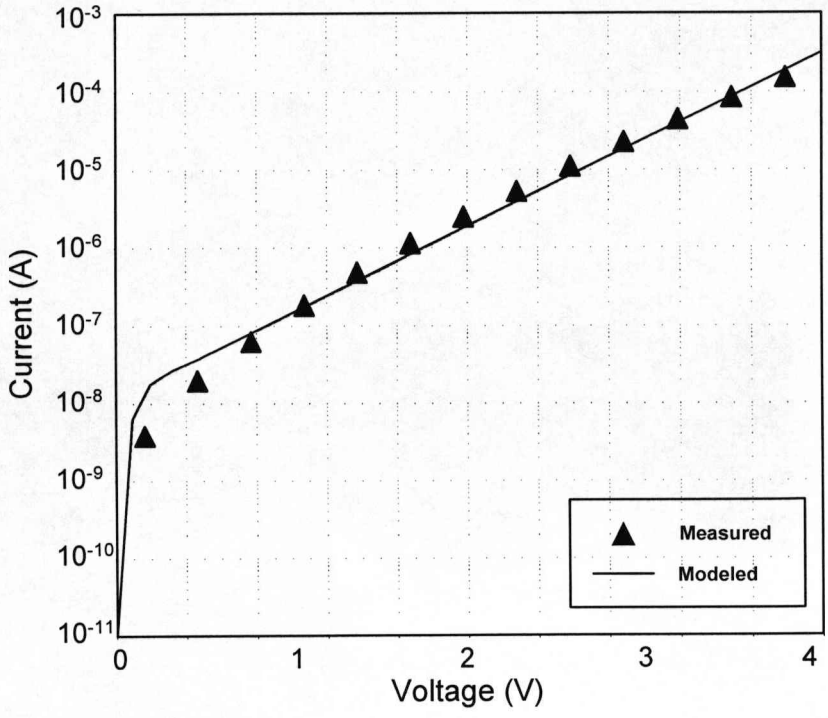


Fig. 5.10 Measured and modeled I-V curves of the reverse-biased zener diode with breakdown voltage of 4.7V.

Section 5.4.1 Synchronous Synaptic Signal

The membrane potential V_{PSP} and the neuron output V_o following the application of a series of pre-synaptic signals with 10ms ISI are shown in Fig. 5.11. In this case all 100 synapses are activated concurrently. The relaxation process of each synapse is controlled by the program voltage of -0.3V enabling the weight charge packet to be generated within 10ms. Therefore each time the pre-synaptic spike arrives, the synaptic array will transmit the same amount of charge to the neuron cell, having the same influence on the membrane potential as demonstrated in Fig. 5.11. Since the membrane potential rises up quickly due to synaptic charge release and then beyond the switching threshold, the CMOS inverters are triggered to produce a significant change in output, the neuron is then said to have fired. It takes about 10ms for the charge to leak away through the reverse-biased zener diode effectively mimicking the repolarising process in the biological neuron cells.

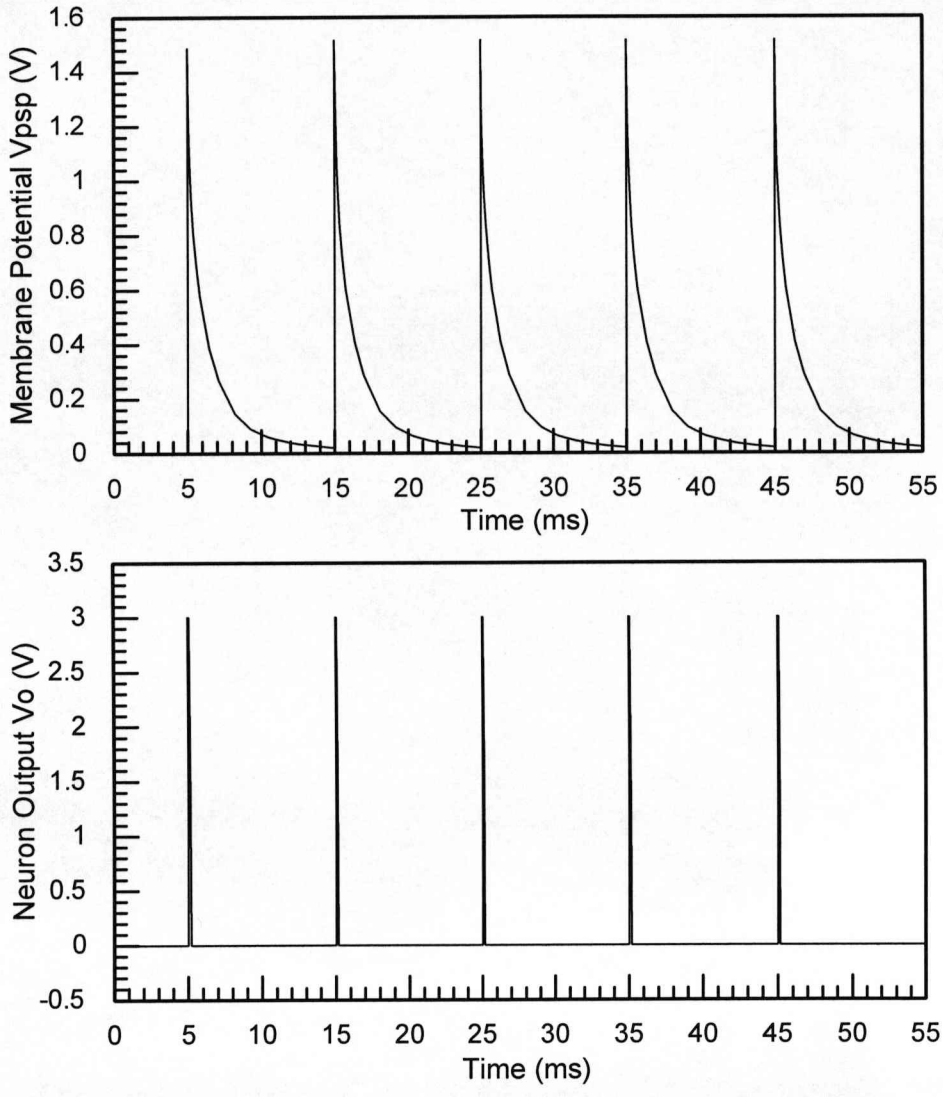


Fig. 5.11 Membrane potential and neuron output in response to the signal from an array of synapses, which are activated at the same time. The pre-synaptic spike arrives at the synapse array at 5ms, 15ms, 25ms, 35ms, and 45ms, with the ISI of 10ms. The weight charge packet in each synapse can be re-generated within 10ms due to the charge injection (V_p=-0.3V).

The response of the neuron cell circuit, for the case of 1ms inter-spike interval (ISI), is shown in Fig. 5.12. The first membrane potential spike, which is large enough to trigger the neuron to fire, reflects the weight charge packet in the full storage well of the synapses. Since the storage well of each synapse can not be refilled within 1ms, the subsequent small amount of charge Q_W from the synapses is unable to charge the membrane node to the threshold required to fire the neuron, exhibiting the depressing behavior as shown in the figure.

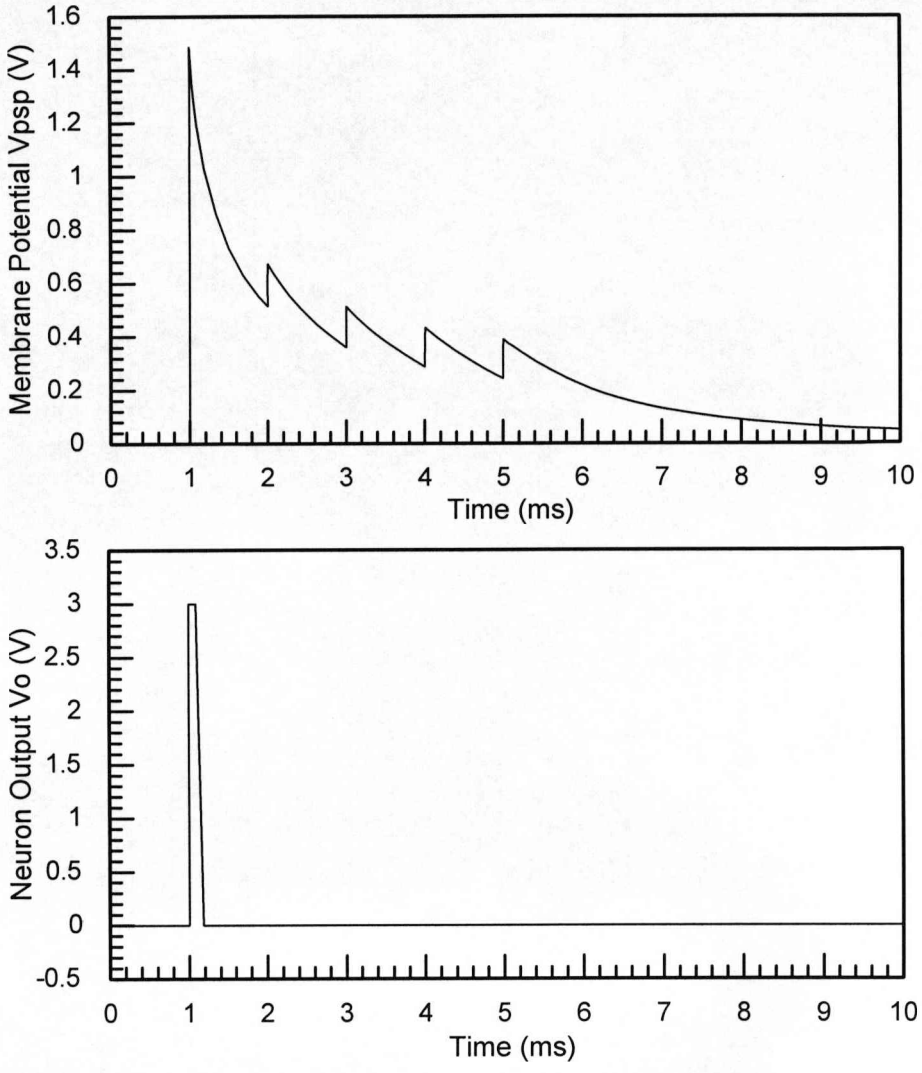


Fig. 5.12 Membrane potential and neuron output in response to the signal from an array of synapses activated at the same time. The pre-synaptic spike arrives at the synapse array at 1ms, 2ms, 3ms, 4ms, and 5ms, with the ISI of 1ms, which is too short to allow the weight charge packet to be fully re-generated in the storage well in each synapse.

Section 5.4.2 Asynchronous Synaptic Signal

Fig. 5.13 shows the membrane potential for a number of sequential spikes from five synaptic groups, each of which has 20 synapses, in response to two pre-synaptic spikes with 10ms ISI. With 0.1ms time lag, the spikes emitted by the five groups of synapses are accumulated at the membrane node, causing the potential to increase

gradually. As soon as the threshold is reached, the neuron fires and sends out a spike, as shown in Fig. 5.13. Due to the charge injection, the storage well of each synapse has been fully refilled when the second pre-synaptic spike arrives at the synaptic array. Therefore the same spiking behavior is observed at the membrane node and neuron output node. Note that the interval between five groups of synapses is chosen such that the accumulation will not be inhibited by the charge leakage, which takes the time of the order of milliseconds.

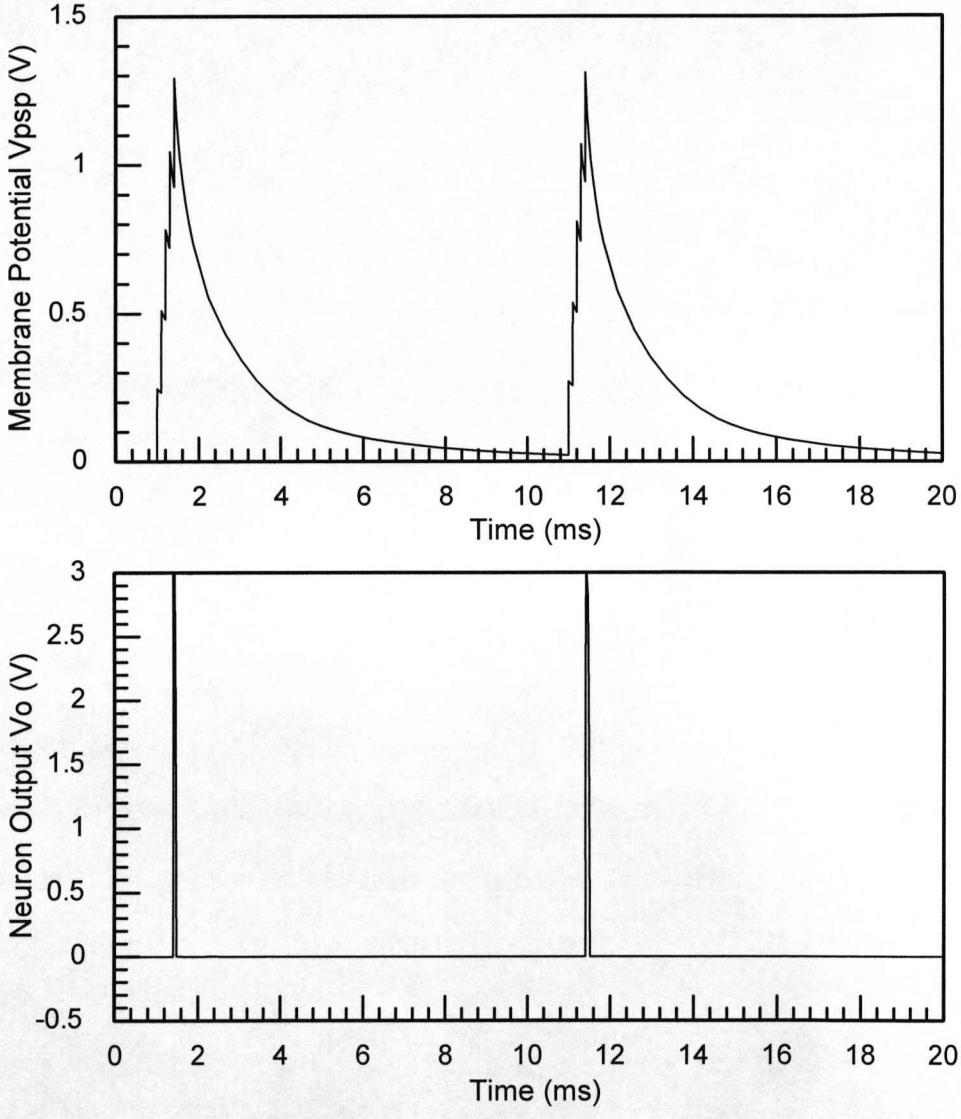


Fig. 5.13 Membrane potential and neuron output against time for sequential facilitating signal from 5 groups of synapses with 0.1ms time lag. The ISI of two pre-synaptic signals is 10ms. The weight charge packet in each synapse can be re-generated within 10ms due to the charge injection (V_p=-0.3V).

Finally, the membrane potential, induced by the spikes from five groups of synapses in responses to two pre-synaptic spikes with 1ms ISI, is shown in Fig. 5.14. After the firing of the neuron caused by accumulative signal from five groups of synapses, the leaky diode conducts a large current. The synapses are then inhibited since the ISI of 1ms is too short to allow the synapse to fully recover the storage well. Therefore the upcoming charge emitted by the inhibitory synapses is unable to balance the leaky charge. Such a small amount of charge has a negligible effect on the firing activity of the neuron.

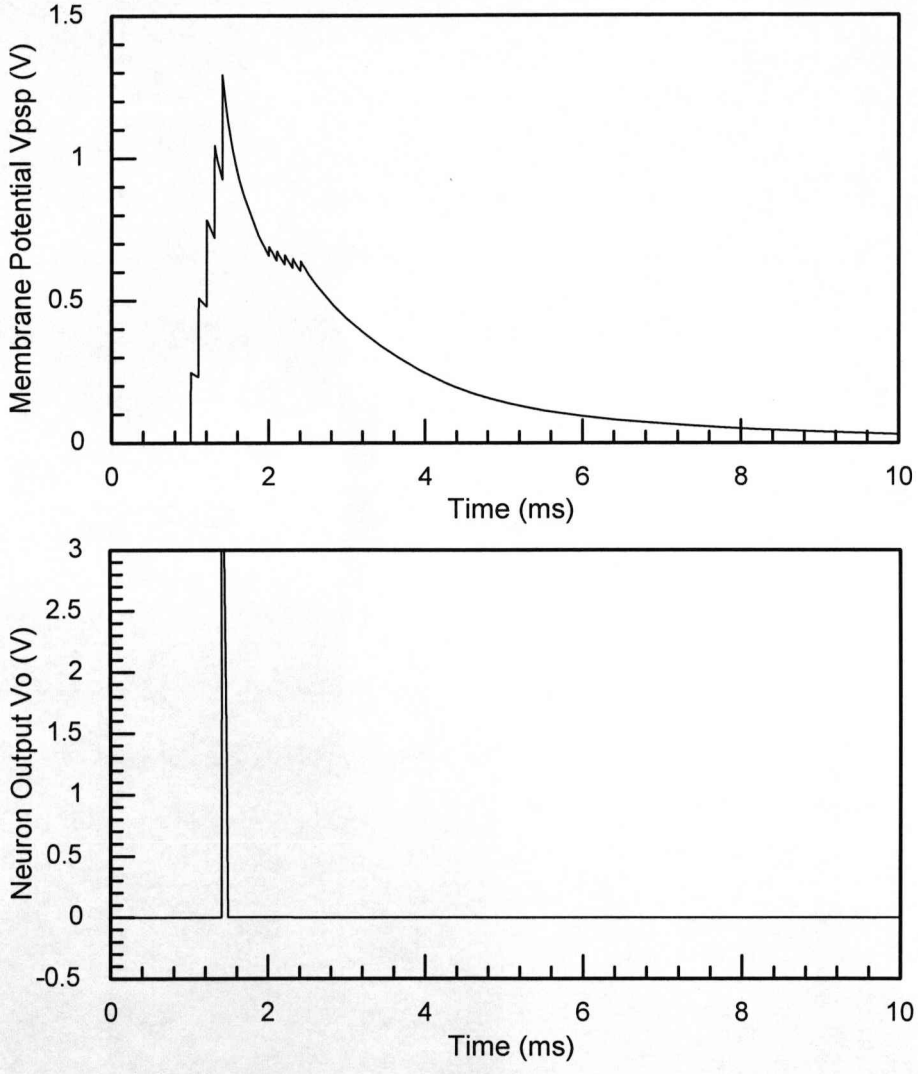


Fig. 5.14 Membrane potential and neuron output against time for sequential depressing signal from 5 groups of synapses with 0.1ms time lag. The ISI of two pre-synaptic signals is 1ms, which is too short to allow the weight charge packet to be fully re-generated in the storage well in each synapse.

Section 5.5 Integration of Synapses with Injector Transistor

In this section an array of programmable dynamic synapses with the modified neuron cell circuit is considered and developed, as shown in Fig. 5.15. Referring to Section 4.5, successive spikes from the pre-synaptic neuron are applied to the gate of the MOS capacitor C2 and repeatedly empty the well under C1, of charge. After each spike event, the well is filled by charge delivered from the injector MOS transistor Mi, which operates in the subthreshold regime. The magnitude of this subthreshold current, which is controlled by the gate voltage, sets the minimum ISI associated with the pre-synaptic train. As before, the neuron cell circuit consists of a current mirror configuration but rather than a leaky diode, a leaky transistor M7, operating in subthreshold, is used. This removes process dependence and adds further programmable control to the cell via V_{Tune}. Current spikes from the array are summed in the drain terminal of M1 and mirrored in the drain terminal of M2. Therefore, the total current, representing the integration of temporal and spatial spikes from n synapses, charges the membrane node potential V_{PSP}. However, this node slowly discharges due to leakage by subthreshold conduction in M7, showing a characteristic latency-intensity function of biological neurons where there is a progressive decrease in latency of membrane potential with increasing input strength. Moreover, by varying the gate voltage, V_{Tune} of this transistor, the voltage decay time constant associated with V_{PSP} is altered: the voltage decay rate at this node is dependant on the capacitance and voltage at the node and the magnitude of the subthreshold current in M7. Note that V_{PSP} can also be used to drive existing conductance-based and phenomenological point neurons [5].

The neuron cell circuit of Fig. 5.15 was simulated in PSpice using the AMS 0.35μm CMOS mixed-signal process. The threshold voltages of the p- and n-channel transistors are given in Table 5.1. For all the transistors, the aspect ratio is unity. The voltage V_{Tune} was set to 0.2V and V_{dd} was set to 3V. An array of two synapses was considered to be integrated into the neuron cell circuit. The program voltage V_{p} of each synapse was set to 0.3V enabling the synapse to re-establish the weight charge packet in 12ms. The output current spikes of the synapses were modeled by a

piecewise linear current source, representing the dynamics of the synaptic outputs shown in Fig. 4.16.

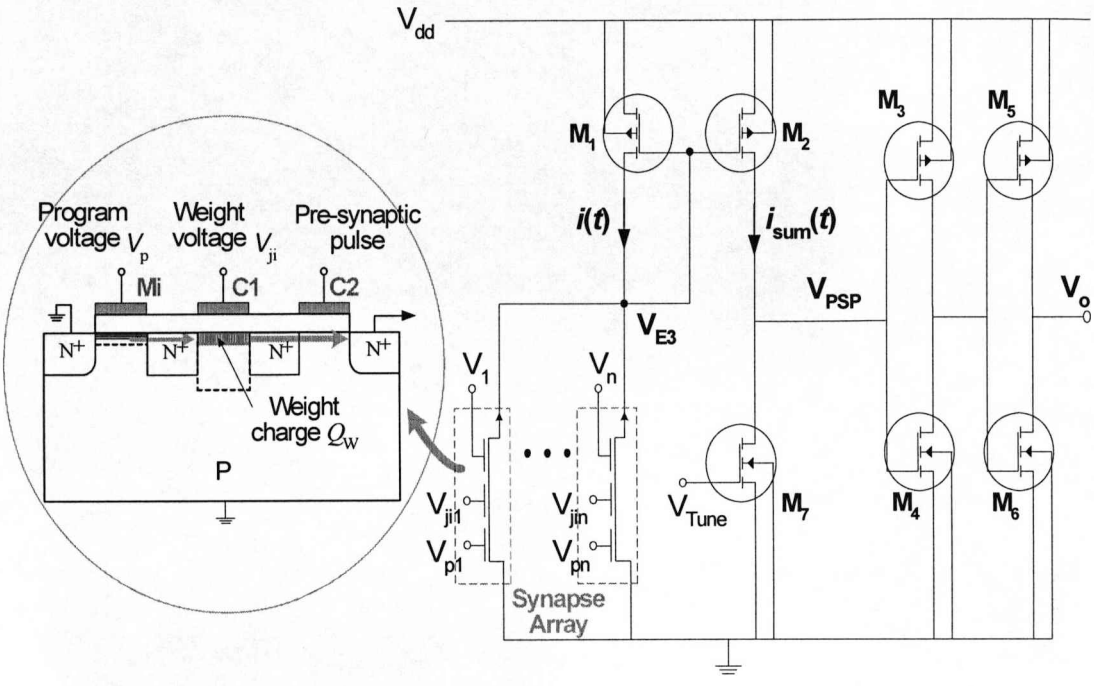


Fig. 5.15 Neuron cell circuit with an array of programmable dynamic synapses. The outputs of *n* synapses are summed in the current mirror circuit which charges the membrane voltage V_{PSP}.

The membrane potential and the output of the neuron cell are shown in Fig. 5.16. The first membrane potential spike increases with the signal from synapse array then reaches the threshold of the neuron. The $CMOS_{5,6}$ inverter makes a lo-hi transition and the neuron is said to have fired. The detailed form of the membrane potential is shown in the inset. The decay of the membrane potential is in 0.1ms, and is tunable by varying the V_{Tune}. For the pre-synaptic spike train with the ISI of 12ms, the synapse array will produce the same output current to the neuron. Therefore the second induced membrane potential reaches the switching threshold and the neuron exhibits the same firing behavior as before.

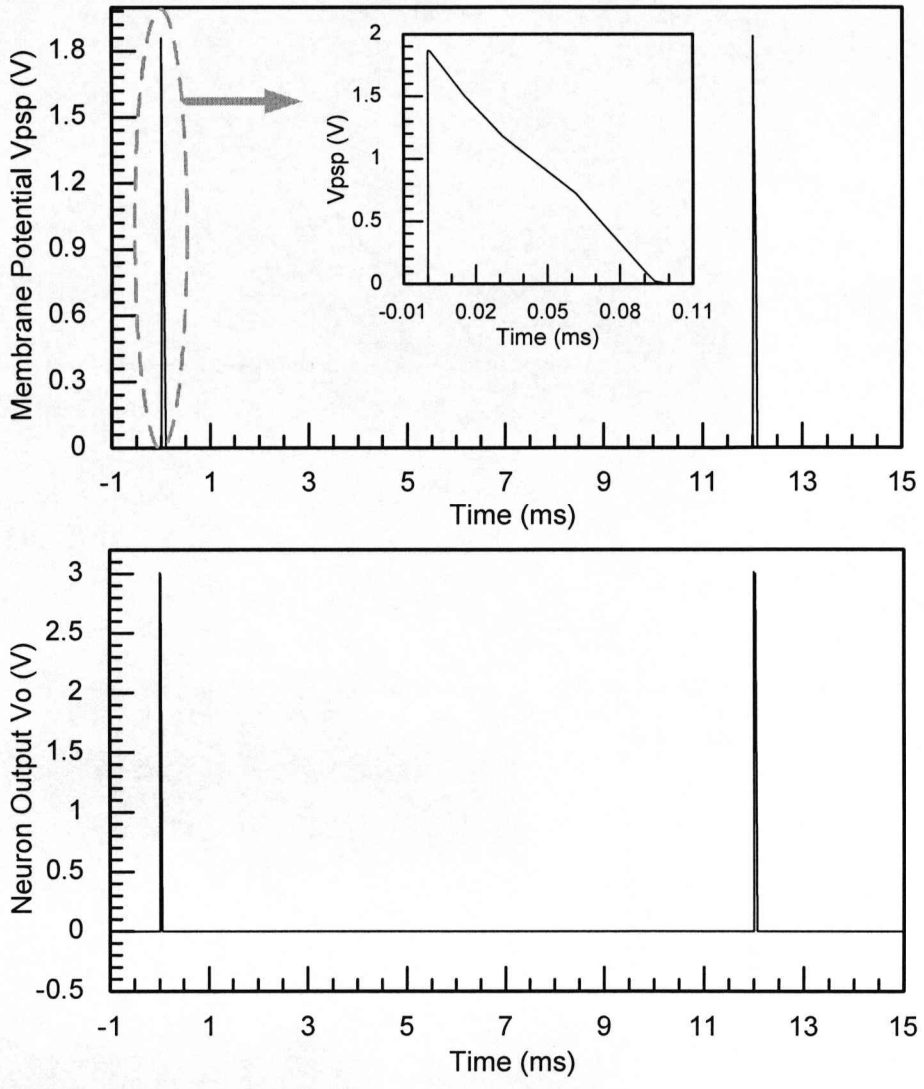


Fig. 5.16 The membrane potential and neuron output for two successive pre-synaptic spikes with 12ms ISI. The program voltage V_p is 0.3V, allowing the synapse to re-establish the weight charge packet in 12ms. V_{Tune} is 0.2V. The inset shows a more detailed view of a membrane spike, which lasts a hundred microseconds.

If the ISI of two successive pre-synaptic spikes is 2ms, the activated synapses will produce a relative low output current as shown in Fig. 4.16. Therefore the first spike of the membrane potential, reflective of the full storage well, will trigger the neuron to fire and decays in 95μs. While the subsequent weak stimuli is unable to manage to get the membrane potential over threshold, inhibiting the neuron to perform the firing

activity as shown in Fig. 5.17. The latency of the second spike of membrane potential is 60μs.

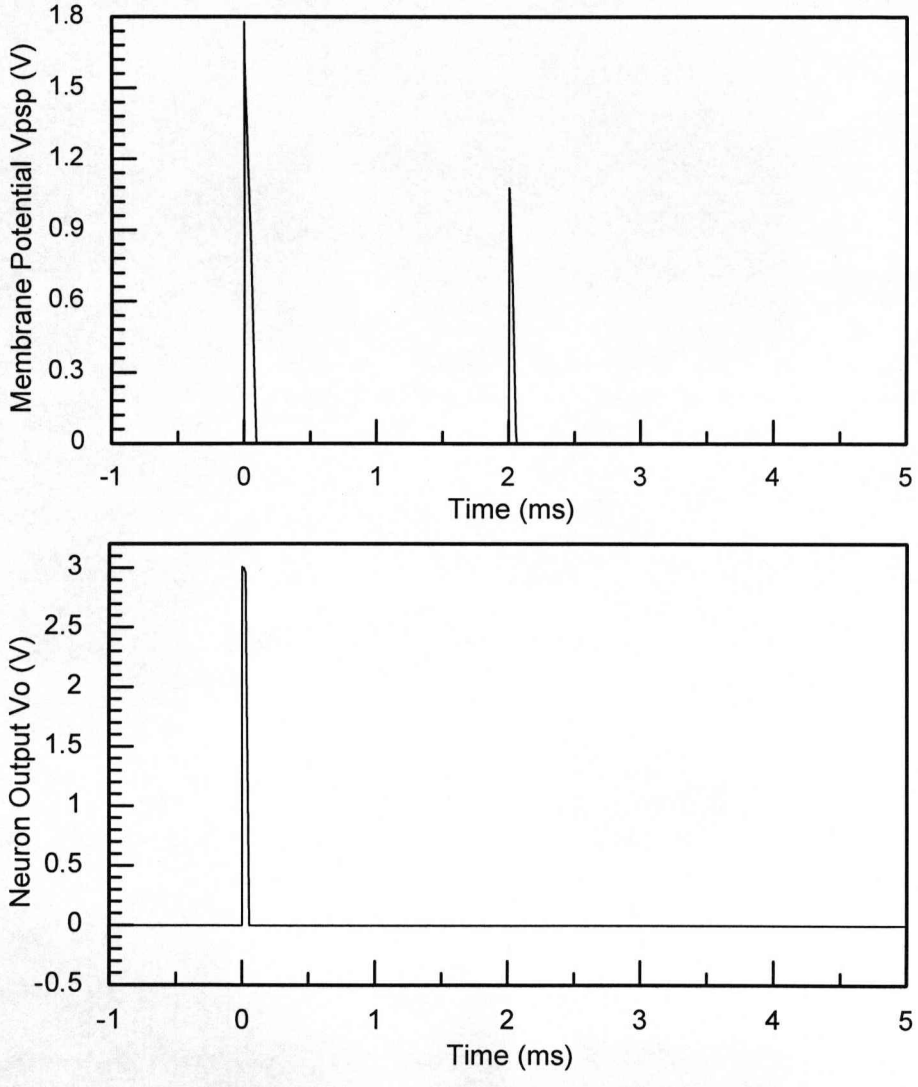


Fig. 5.17 The membrane potential and neuron output for two successive pre-synaptic spikes with 2ms ISI, which is too short to allow the synapse to re-establish the weight charge packet. V_{Tune} is 0.2V.

The facilitation behavior resulting from successive spikes from one activated synapse is demonstrated in Fig. 5.18. In this plot an input spike train of 1MHz, corresponding to an ISI of 1μs, was applied to the drain of M1: the parameters for the simulation are shown in the inset. As soon as the switching threshold of $CMOS_{3,4}$ is reached, the

neuron circuit fires. Note that the transfer of charge from the well in the synapse to the membrane node V_{PSP} can be considered to be instantaneous. However, as the number of synapses is increased the associated capacitance of the N+ diffusion regions also increases and consequently the charge transfer rate will diminish. It can be shown that the number of synapses, or fan-in, is limited by the transfer rate, which correlates with the magnitude of this capacitance. Re-arranging (5.4) shows that the transfer duration is related to the fan-in n and N+ node diffusion capacitance C_{FN} by:

$$\tau_r = \left(\frac{mV_t}{I_0}\right) nC_{FN} \tag{5.18}$$

where m and I_0 are the gate channel coupling and off-current associated with M1 respectively. Clearly the transfer duration τ_r is proportional to n and this relationship will have important implications for scaling.

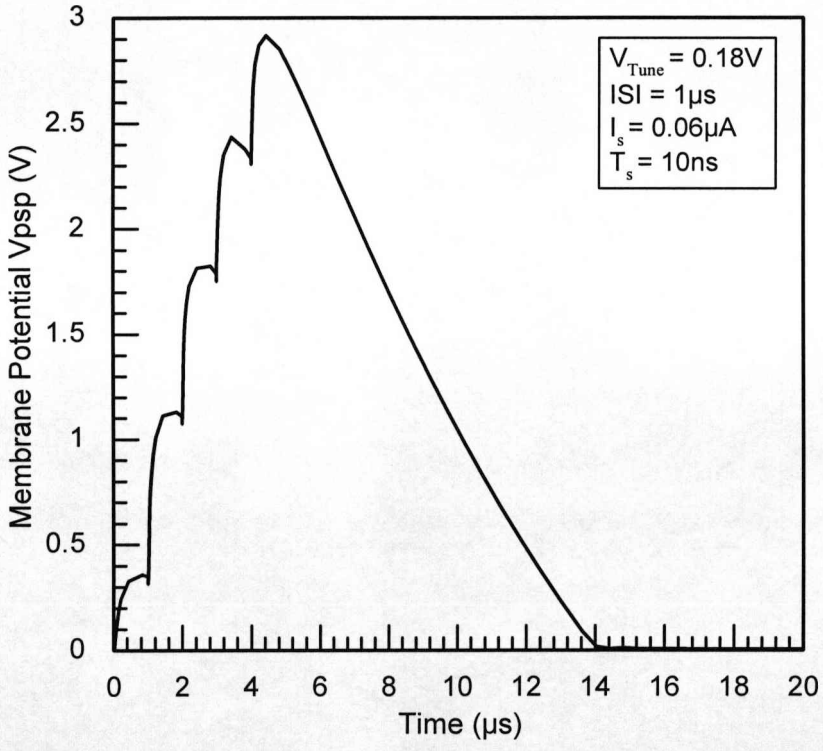


Fig. 5.18 Facilitating response of the membrane potential V_{PSP} to a series of pre-synaptic signals with ISI=1µs.

The simulation results clearly demonstrate the facilitating capability where a simple current mirror configuration can sum temporal/spatial spike distributions and represent the aggregate as a voltage at the membrane node. The significantly different rise and fall time constants commonly associated with the membrane node potential is controlled by using the subthreshold leakage current of a MOS transistor. It is worth noting that in practice, the differing drain voltages on each branch of the current mirror will result in a mismatch of the two drain currents. However this not a concern, since the main function of the current mirror configuration here is to provide a summing action of synaptic outputs and to transfer the associated charge onto the capacitance associated with the membrane node potential.

Section 5.6 Neuron MOS Transistor

In this section an alternative approach for spiking neuron cell implementation is described. The neuron is implemented as a multi-input floating gate MOS transistor (neuMOS) providing the weighted summation and thresholding functions [6]. The charge coupled synapse, presented in Chapter 3, is connected to a sub-gate of a neuMOS via a 'self biasing' floating diffusion region realizing a compact lower power standard neuron cell.

Section 5.6.1 neuMOS Principles

The basic structure of the neuMOS is shown in Fig. 5.19(a). This *n*-channel MOS transistor has multiples of input gates capacitively coupled to a common floating gate. The input voltages and coupling coefficients are illustrated in Fig. 5.19(b), where V_1, V_2, . . . , V_n are the input voltages; C_1, C_2, . . . , C_n are the coupling capacitance between the floating gate and each input gates; C_0 is the coefficient between the floating gate and the substrate; V_F is the floating gate voltage.

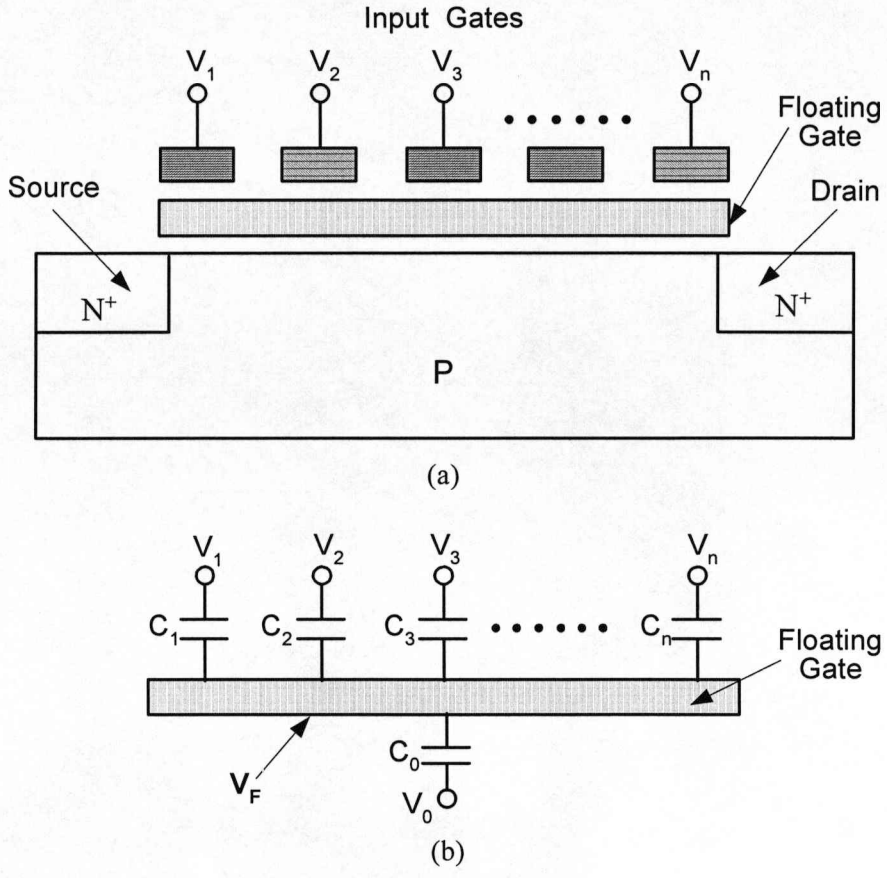


Fig. 5.19 (a) Schematic drawing of the neuMOS structure; (b) Capacitive model of the neuMOS.

The initial net charge on the floating gate is assumed to be zero, and it is noted that no actual charge injection occurs during the device operation, rather the floating gate potential is modulated by capacitive coupling from the input gates. The substrate and source are grounded. Then the floating gate potential is:

$$V_F = \frac{C_1 V_1 + C_2 V_2 + \cdots + C_n V_n}{C_T} \tag{5.19}$$

where C_T is the total capacitance associated with the floating gate, which is given by:

$$C_T = \sum_{i=0}^{n} C_i \tag{5.20}$$

This is the most important feature of the neuMOS, which states that the floating gate potential is determined as a linear sum of all input signals weighted by the capacitive

coupling coefficients [6]. The formation of the channel underneath the oxide of the neuMOS is controlled by this gate level sum operation. Thus the 'on' and 'off' of the transistor depend on the result of weighted sum operation. The transistor turns on at the condition:

$$V_F = \frac{C_1V_1 + C_2V_2 + \cdots + C_nV_n}{C_T} > V_T \tag{5.21}$$

where V_T is the threshold voltage of the transistor as seen from the floating gate.

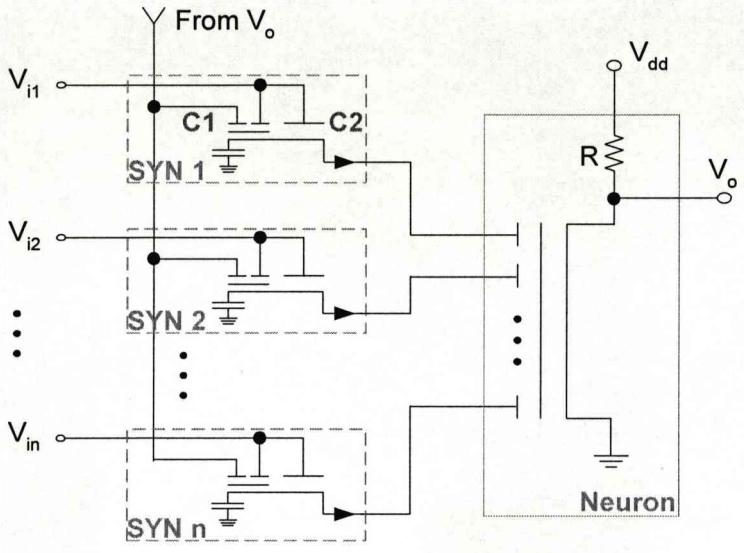
This threshold operation of the transistor resembles the behavior of a biological neuron if the turn-on of the transistor is correlated to the firing of a neuron. Since the weighted summation is performed in a voltage mode utilizing the coupling effect, essentially no power dissipation occurs in the calculation, making the device ideal for VLSI implementation. Note that application of substrate bias to the neuMOS can provide an additional tuning capability for the neuron threshold level.

Section 5.6.2 Interconnecting Regime

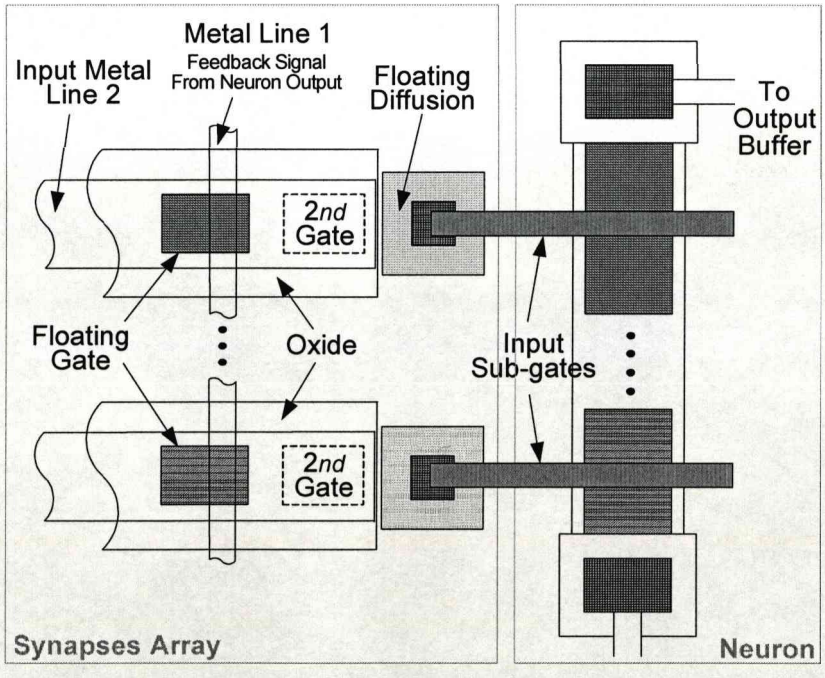
The charge coupled synapses can be integrated with the neuMOS to form a highly compact standard neuron cell, as illustrated in Fig. 5.20 where the schematic and top view of the design architecture are shown. Each synaptic output is assigned to one of the sub-gate contacts and the synaptic array with common floating diffusion outputs can be readily integrated with the neuMOS. Therefore the output of the charge coupled synapse is read directly by the point neuron. Due to the capacitive coupling to the floating gate of neuMOS, the effective charge transfer to the floating gate represents the aggregate of the total charge transferred from all the silicon synapses. After the signal is captured by the neuMOS, the floating diffusion and MOS capacitors of the synapse relax to equilibrium by thermal generation of charge, mimicking the decay process of PSP in real synaptic operation.

As indicated in Fig. 5.20, contact is made by the input line to both the gate of C2 and the floating gate of C1. Feedback of PSPs from the neuron output to C1 is used, in conjunction with the signal on the input line, to update the weight voltage V_{ji} stored in

the floating gate as well although the physical details of this updating operation are not considered here. The necessary conditions on the two connections to the first MOS capacitor C1 will enable the localized training dynamics of the neuron cell using the Hebbian learning rule. Memory cells controlled by two input control signals have been described previously [7] so engineering solutions are available.



(a)



(b)

Fig. 5.20 (a) Schematic of the basic silicon neuron cell consisting of a neuMOS associated with the charge coupled synapses; (b) Top view of the layout.

Section 5.7 Discussion and Conclusions

In this chapter, the development of different standard neuron cells comprising a point neuron and its associated synapses has been presented. The output terminal of the charge coupled synapse constitutes the interface with the point neuron. The fundamental functionality of the biological neuron cell is implemented by current mirror summing, charge integration onto a thresholding inverter and subsequent slow leakage of charge via a reverse-biased diode or an MOS transistor in subthreshold. The simulations and measurement on the spiking neuron circuit with different synapse types have demonstrated a number of key features of the biological neuron cell. This simple spiking neuron circuit possesses the potential of large scale implementation of massively interconnected neural networks. An alternative approach for the summing and thresholding operations has been considered in this chapter, based on a neuMOS. Synapses can be assigned to each of the neuMOS inputs. The approach is feasible but there are some challenges in implementing it due to the nature of the signal generated at the output which requires additional circuitry to de-embed the information.

References

[1] W. Gerstner, W.M. Kistler, *Spiking Neuron Models: Single Neurons, Populations, Plasticity*, Cambridge University Press, 2002.

[2] C. A. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.

[3] S. R. Campbell, D. L. Wang, and C. Jayaprakash, "Synchrony and desynchrony in integrate-and-fire oscillators," *Neural Comput.*, vol. 11, pp. 1595–1619, 1999.

[4] J. V. Arthur and K. Boahen, "Synchrony in silicon: the gamma rhythm," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1815–1825, 2007.

[5] E. M. Izhikevich, "Which model to use for cortical spiking neurons?" *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1063–1070, 2004.

[6] T. Shibata and T. Ohmi, "A functional MOS transistor featuring gate level weighted sum and threshold operations," *IEEE Transactions on Electron Devices*, vol. 39, no. 6, pp. 1444–1455, June 1992.

[7] K. Hieda, M. Wada, T. Shibata, S. Inoue, M. Momodomi, and H. Iizuka, "Optimum design of dual-control gate cell for high-density EEPROM's," *IEEE Transactions on Electron Devices*, vol. ED-32, no. 9, pp. 1776–1780, Sep. 1985.

CHAPTER 6 CONCLUSIONS AND FUTURE WORK

In the last few decades, much significant research has been carried out on the development and deployment of engineering neural network models that can be implemented in both hardware and software and used to inspire new paradigms for real time computational networks. However, neural systems are difficult to model as they consist of many nonlinear elements in massively parallel and have a substantial range of time constants. The neural behavior cannot be solved analytically, and the speed of the simulation is limited by longest time constant, even for a modern fast computer system. By contrast, physical devices such as semiconductors enable the neural system models to operate in real time, and make the speed of the system models independent of the parallelism associated with these systems. This project seeks to develop a compact hardware implementation of synapses that, in so far as possible, capture the intrinsic dynamics of real synapses which are the dominant building block in neural networks. In particular, the project explores the potential of a charge transfer device structure as a synaptic node and a simple circuit as a point neuron. The silicon synapse and neuron are constrained to operate under strict low power conditions, and offer the potential of scaling to massively parallel networks specific to spike-based neural systems. The achievements of the work presented in this thesis are summarized in this chapter, with suggestions for future work.

In Chapter 1, a brief review of the fundamentals of neural networks is presented. The so-called spiking neural networks and its neuron models take the advantages over the conventional artificial neural networks, and exhibit the synaptic plasticity which is a form of change of the pre-processing, that plays an important role in the learning of neural networks. This chapter also reviews the state of the art in conventional neural hardware, outlines the fundamental principles and concepts of implementation technology, and presents existing work on electronic implementation of neural networks, especially the basic units: neurons and synapses. Using VLSI technology, such silicon models of neurons together with the models of synapses could prove useful in robots, sensors, and many other research areas.

Chapter 2 presents some relevant fundamental aspects of semiconductor devices. The explanation of three basic operation modes of a MOS capacitor, accumulation, depletion and inversion, is provided. The C-V characteristics of the MOS capacitors are described and illustrated by an experimental study. In addition, the principles of the MOS transistor action are provided. The semiconductor industry has been one of the most successful on the planet. Moore's law has been the guiding principal for semiconductor industry over the last 40 years. Until now, the regular increase in the computational and processing capability of integrated circuits has resulted from making smaller silicon-based transistors. As CMOS technology starts to run into its fundamental limit, there has been an interest in exploring bio-inspired architectures which is the focus of this work.

In Chapter 3, we have developed a charge coupled synapse, based on a two-capacitor charge transfer device structure. The weighting functionality can be integrated into the first stage by means of a floating gate. A pre-synaptic spike to the second phase allows the charge under the first gate to drift onto the output terminal, which constitutes the interface with the point neuron, to produce a current or voltage spike. The simulation results show that the implementation of synapse represents the intrinsic dynamics of biological synapse, the dominant building block in spiking neural networks, by using innate features of the semiconductor physics. Since most of the operations are capacitive in nature, the proposed implementation is inherently low power and simple calculations indicate that the new device can surpass the performance of current implementation techniques.

In Chapter 4, the establishment of weight charge packet by thermal generation of electron-hole pairs is mathematically modeled, and the response of a charge coupled synapse to successive pre-synaptic spikes is investigated. To speed up the recovery process of charge coupled synapses and make the time scale of the ISI of biological system tunable, we proposed two implementations of programmable dynamic synapses integrated within single semiconductor devices. The first is implemented using the charge coupled synapse structure to which an additional source of minority carriers is attached. The programmable functionality of the synapse is realized by the weight restoration through charge injection from an n^{+}-p diode pulsed by a small negative voltage. The second comprises a MOS transistor operating in subthreshold

and two MOS capacitors in proximity to the transistor. One of the capacitors is permanently biased in strong inversion where the associated density of charge in the well implements the weighting. When a pre-synaptic spike is applied to the gate of the second MOS capacitor the charge density in the well falls producing a current spike at the output. The amplitude of the spike is correlated with the equilibrium charge density in the well, which is controlled by the associated gate voltage. The function of the MOS transistor is to restore the charge in the well whereby the duration of this process is dictated by the associated gate voltage. Therefore, the synapse is capable of operating in the facilitating state over a large frequency range. The area of a programmable dynamic synapse can be maintained to be small and since it operates in transient mode, its power consumption is negligible. Simulation results are presented which clearly demonstrate its operation.

Chapter 5 presented an analog neuron cell with the newly developed charge coupled synapses. Aggregation of spikes from an array of synapses is achieved using a current mirror configuration whose output post-synaptic potential can be used to stimulate a point neuron circuit. A couple of CMOS inverters were employed to implement the temporal and spatial integration of weighted input spikes generated by the synaptic array. The decay of the membrane potential is mimicked by the charge leakage through a reverse-biased diode, or a tunable MOS transistor. The neuron cell is capable of capturing the summing and thresholding dynamics of biological neurons. Simulations were presented to verify that the proposed neuron cell implementation is feasible and has the potential for implementing biological neural networks in hardware.

There are three aspects of the future work on the basis of the work in this thesis. For the implementation of dynamic synapses, the further work involves extending the dynamic synapse, presented in Section 4.5, to operate in the depressing state by controlling the rate at which the potential well empties. Both simulation and experimental results from test structures should be presented. This work will facilitate the restoration of weight charge packet more dynamical and programmable.

The second is related to the floating gate technology. As mentioned above, the synaptic weight is represented by the stored charge with associated voltage on the

floating gate where the charge would be added or removed through a tunneling process using well-established, read-only memory cell technology. The synaptic weight can therefore be updated by engineering the correlation between the pre-synaptic and the post-synaptic signals. In such memory devices, there are essentially no pathways through which charge can flow, because the conductive element is completely surrounded by dielectric making the capacitor isolated from all other circuits. The threshold voltage can be altered by changing the amount of charge present between the channel and the gate. The state can be set by using both Fowler-Nordheim tunneling and hot electron injection, which require high voltages and a relatively long time. In general, a feedback is required to improve the accuracy and precision of the floating gate devices. On architectural level neural implementation, issues such as the choice of training algorithms, network size, application domain, which will influence the choice of floating gate technology, need to be considered.

In the human brain, neurons of the order of 10^{11} are interconnected in a complex pattern, it is inconceivable that current metal layout topologies will facilitate this level of integration primarily because the silicon surface area consumed by interconnect is proportional to the product of the neuron density in adjacent layers. Therefore the inter-neuron connectivity problem should be addressed. In particular, a time multiplexing neural network architecture, whereby the interconnect density is independent of the number of neurons in the network, should be explored to provide a very area efficient route to interconnect implementation. In addition, there is a need to design low power control circuitry for the neurons and synapses at system level, where it may be possible to use, for instance, subthreshold logic, given the overall slow speed of the circuitry. Furthermore, the potential to scale the devices to the densities required for emulating brain-like behavior should be explored. The entire silicon neural network with the compact low power analog neuron cells will advance both the computational intelligence and the CMOS technology well beyond its current "post CMOS scaling limit".

Appendix 1 Description of Spice Model Parameters

Parameter	Description
W	Channel width
L	Channel length
LEVEL	Model index
LD	Lateral diffusion
TOX	Oxide thickness
VTO	Zero-bias threshold voltage
KP	Transconductance parameter
NSUB	Substrate doping
GAMMA	Bulk threshold parameter
PHI	Surface potential
UO	Surface mobility
UEXP	Critical field exponent in mobility degradation
UCRIT	Critical field for mobility degradation
DELTA	Width effect on threshold voltage
VMAX	Maximum drift velocity of carriers
XJ	Metallurgical junction depth
LAMBDA	Channel length modulation
NFS	Fast surface state density
NEFF	Total channel charge (fixed and mobile) coefficient
NSS	Surface state density
RSH	Drain and source diffusion sheet resistance
PB	Bulk junction potential
CGDO	Gate-drain overlap capacitance, per channel width
CGSO	Gate-source overlap capacitance, per channel width
CJ	Zero-bias bulk junction bottom capacitance per m^2 of junction area
MJ	Bulk junction bottom grading coefficient
CJSW	Zero-bias bulk junction sidewall capacitance per m of junction perimeter
MJSW	Bulk junction sidewall grading coefficient

Appendix 2 Associated Publications

[1] Y. Chen, L. McDaid, S. Hall, and P. Kelly, "A programmable facilitating synapse device", *Proceeding of the 2008 IEEE World Congress on Computational Intelligence (A joint conference of the IJCNN, FUZZ-IEEE and CEC)*, pp. 1615–1620, Hongkong, June 2008.

[2] L. McDaid, J. Harkin, S. Hall, T. Dowrick, Y. Chen, and J. Marsland, "EMBRACE: Emulating biologically–inspired architectures on hardware", (Invited paper) *9th WSEAS International Conference on Neural Networks*, Sofia, Bulgaria, May 2008.

[3] L. McDaid, S. Hall, Y. Chen, O. Buiu, and P. Kelly, "A biologically plausible spiking neuron cell in hardware," *IEEE Transactions on Neural Networks*, submitted.

[4] Y. Chen, S. Hall, L. McDaid, O. Buiu, and P. Kelly, "A solid-state neuron for spiking neural network implementation", *Engineering Letters*, vol. 16, no. 1, pp. 83–89, 2008.

[5] Y. Chen, S. Hall, L. McDaid, O. Buiu, and P. Kelly, "Analog spiking neuron with charge-coupled synapses", *Proceedings of the World Congress on Engineering 2007*, London, UK, vol. 1, pp. 440–444, July 2007.

[6] Y. Chen, S. Hall, L. McDaid, O. Buiu, and P. Kelly, "A solid state neuron for the realisation of highly scaleable third generation neural networks", *Proceeding of the 2006 International Conference on Solid-State and Integrated Circuit Technology*, pp. 1071–1073, Oct. 2006.

[7] Y. Chen, S. Hall, L. McDaid, O. Buiu, and P. Kelly, "On the design of a low power compact spiking neuron cell based on charge-coupled synapses", *Proceeding of the 2006 IEEE World Congress on Computational Intelligence (A joint conference of the IJCNN, FUZZ-IEEE and CEC)*, pp. 1511–1517, Vancouver, Canada, July 2006.

[8] Y. Chen, S. Hall, L. McDaid, O. Buiu, and P. Kelly, "A silicon synapse based on a charge transfer device for spiking neural network application", *Lecture Notes in Computer Science*, vol. 3973, pp. 1366–1373, Berlin: Springer-Verlag, 2006.

[9] Y. Chen, S. Hall, L. McDaid, and P. Kelly, "Silicon synapse for hebbian learning application", *Proceeding of EPSRC PREP'05*, pp. 71–72, Lancaster, UK, March 2005.